Validating Market Risk Models Using Realized PIT Values

Lok Hsiao Yen

PhD in Actuarial Science

Heriot Watt University, Edinburgh

School of Mathematical and Computer Sciences

November 2017

Abstract

The aim of this thesis is to propose new tests for validating market risk models for financial losses using realized probability-integral-transform (PIT) values. We introduce a flexible framework for testing Value-at-Risk (VaR) exceptions at multiple levels based on a weighted transformation of the realized PIT values, where the weight function reflects the risk objectives of the modeller. This framework can be extended to perform tests using multiple different weight functions. We show that this extended framework either nests or is closely related to many of the traditional VaR and realized PIT tests in existing literature. This approach to model validation is preferable to likelihood-ratio based testing, which can be shown to be a test that is based on a set of specific weight functions that may not reflect the modeller's risk objectives. A further advantage of this framework is that it can be easily be extended to explicitly tests for serial independence of the realized PIT values. We do this using the idea of blocking, as well as exploiting the martingale difference (MD) property. In the empirical studies, we have also found that tests based on VaR exceptions have great difficulty in detecting poorly calibrated historical simulation (HS) models. We will see how tests based on elicitability theory and proper scoring rule complement the tests based on VaR exceptions to detect poorly calibrated HS models.

Acknowledgement

I am very grateful for the many inputs and advice from my supervisor, Alexander J. McNeil, during the course of the PhD, which are crucial for the success of this research project. I am also grateful for the useful inputs from my co-authors Marie Kratz (Kratz et al., 2016) and Michael Gordy (Gordy et al., 2017). I acknowledge that this research project is funded by the Actuarial Research Center and the Institute and Faculty of Actuaries.

ACADEMIC REGISTRY
Research Thesis Submission

Name*:*	LOK HSIAO YEN		
School:	Mathematical and Computer Sciences		
Version: *(i.e. First, Resubmission, Final)*	First	Degree Sought:	PhD in Actuarial Science

Declaration

In accordance with the appropriate regulations I hereby submit my thesis and I declare that:

1) the thesis embodies the results of my own work and has been composed by myself
2) where appropriate, I have made acknowledgement of the work of others and have made reference to work carried out in collaboration with other persons
3) the thesis is the correct version of the thesis for submission and is the same version as any electronic versions submitted*.
4) my thesis for the award referred to, deposited in the Heriot-Watt University Library, should be made available for loan or photocopying and be available via the Institutional Repository, subject to such conditions as the Librarian may require
5) I understand that as a student of the University I am required to abide by the Regulations of the University and to conform to its discipline.
6) I confirm that the thesis has been verified against plagiarism via an approved plagiarism detection application e.g. Turnitin.

* *Please note that it is the responsibility of the candidate to ensure that the correct version of the thesis is submitted.*

Signature of Candidate*:*		Date:	

Submission

Submitted By *(name in capitals):*	LOK HSIAO YEN
Signature of Individual Submitting:	
Date Submitted:	

For Completion in the Student Service Centre (SSC)

Received in the SSC by *(name in capitals):*			
Method of Submission *(Handed in to SSC; posted through internal/external mail):*			
*E-thesis Submitted (**mandatory for final theses**)*			
Signature:		Date:	

Please note this form should be bound into the submitted thesis.
Academic Registry/Version (1) August 2016

Contents

Chapter 1 Introduction

1.1 Aims of the thesis

The aim of this thesis is to propose new tests for validating market risk models for financial losses using realized probability-integral-transform (PIT) values. Traditionally, backtests are mostly based on VaR exceptions. Since realized PIT values contain information of VaR exceptions at any levels, a well-designed test based on realized PIT values should be more powerful than traditional VaR exception-based tests.

We introduce a flexible framework for testing VaR exceptions at multiple levels based on a weighted transformation of the realized PIT values, where the weight function reflects the risk objectives of the modeller. For example, if we are interested in testing the 99% Value-at-Risk (VaR) of a forecast model, we can use a symmetric weight function centered at the 99% level. By doing so, we hope to obtain a more powerful and stable test at the cost of specificity to the risk objective. This framework can be extended to perform tests using multiple different weight functions. By doing so, we test different aspects of the realized PIT values simultaneously, and we show empirically that this improves the power of the test significantly.

We show that this extended framework either nests or is closely related to many of the traditional VaR and realized PIT tests in existing literature. This approach to model validation is preferable to likelihood-ratio based testing, which can be shown to be a test that is based on a set of specific weight functions that may not reflect the modeller's risk objectives. For regulatory purposes, it may be important to use a one-sided test, since regulators are less concerned with banks holding more capital than necessary. We will show how to construct a one-sided test using our framework.

The realized PIT values of a good forecast model should be serially independent. A further advantage of this framework is that it can be easily be extended to explicitly tests for serial independence of the realized PIT values. We do this using the idea of blocking, as well as exploiting the martingale difference (MD) property. Rather than using traditional tests based on auto-correlation function (ACF), the blocking

approach improves the size of the tests at the cost of power, which can be important when we are interested in testing the tail of the forecast models using heavily truncated realized PIT values. Both ACF-based tests and MD-based tests have rather poor size in such cases. For the MD-based tests, we will also show how to perform size correction. The MD approach is more flexible, as it allows the use of factors other than the weighted transformation of the realized PIT values to test for serial independence, which may improve the power of the tests.

Our empirical studies show that by using the absolute realized PIT values as the factor, we are able to significantly improve the ability of the tests to detect serial dependence. In the empirical studies, we have also found that tests based on VaR exceptions have great difficulty in detecting poorly calibrated historical simulation (HS) models. We will see how tests based on elicitability theory and proper scoring rule complement the tests based on VaR exceptions, especially for detecting poorly calibrated HS models.

1.2 Background

Due to lack of complete information, financial events can be viewed as random events. From a risk management perspective, financial institutions are concerned with the potential of these financial events leading to financial losses, and the uncertainty associated with these financial losses, and most importantly, whether the extent of the financial losses will cause the financial institutions to become insolvent. To reduce the risk of insolvency to an acceptable level defined by the regulators, financial institutions are required to demonstrate that they have sufficient capital to act as a buffer.

As the size of the portfolios of financial institutions are typical very large, rather than modeling all possible risk, it is common to focus on the modeling of a smaller subset of risk factors that the portfolios are most sensitive to. The risk factors may include, for example, equity prices, exchange rates, interest rates for different maturities and volatility parameters for valuation models.

We will operate in the probability space $(\Omega, \mathcal{F}, \mathrm{P})$. We denote the realized loss at

time t by L_t, which we will assume to be random at time $t-1$, but can be computed at time t as a function of the risk factor changes in the interval $t-1$ to t, where we assume the composition of the trading book remains fixed in this interval. We denote by $\mathcal{F}_t$ the filtration at time t, which represents the information available to the modeller at time t, and we denote the conditional loss distribution given information up to time $t-1$ by

$$F_t(x) = \mathrm{P}\left(L_t \leqslant x \mid \mathcal{F}_{t-1}\right), \tag{1.1}$$

which we will assume in the sequel to be a continuous distribution for all t. Once a suitable risk model is determined, the risk modeling group will then construct the estimated conditional loss distribution $\widehat{F}_t$ for financial losses based on information up to time $t-1$. We will refer to $\widehat{F}_t$ as the forecast distribution.

Risk measures are mappings of the forecast distribution into real numbers representing the capital amounts required as buffer against insolvency. A risk measure is said to be law-invariant if it depends only on the probability distribution. One popular law-invariant risk measure is Value-at-Risk (VaR). When the forecast distribution is continuous, VaR at level α is the quantile of the forecast distribution at level α. Another popular risk measure is expected shortfall (ES), also known as tail value-at-risk (TVaR). Assuming that the forecast distribution is continuous, the ES at level α is the conditional expected loss given exception of VaR at level α. See Section 3.1 for the formal definition of VaR and ES.

Ideally, we would hope that the risk model chosen by the financial institutions is able to capture the main characteristics of the losses that it is designed to model. However, in practice, it is almost impossible to determine such a risk model. Instead, we should ask whether the estimated risk measures have served its risk management objectives. For example, VaR is chosen as the capital buffer to ensure that the banks are able to survive fairly large unexpected losses. Holding a 99% VaR represents the objective to survive a one in a hundred days loss event, and holding a 99.95% VaR represents the objective to survive a once in a 8-year loss event. The banks however are less interested in extreme losses where they have close to zero probability of survival. Holding sufficient capital for such an objective will have a large impact on profitability of the banks. Hence, the validation of risk models calibrated to have accurate estimates of 99% VaR should focus on the region close to the 99% level.

A backtest performs a statistical test for the null hypothesis that the risk management objectives have been met. See Chapter 2 for more details on the concept of hypothesis testing. In order to perform a backtest, we first need to determine the objectives. For a given objective, we can either evaluate a set of risk measure estimates obtained from the forecast distribution, or the forecast distribution itself, by assigning test input values to the realized losses, observed ex post, based on some test input function. The choice of the test input function should align with the risk management objectives. We then perform tests on the test input values.

For the evaluation of forecast distributions, a popular test input function is the probability-integral-transform (PIT), which is the estimated probability of observing a loss no more extreme than a particular ex post loss, computed using the forecast distribution. We will subsequently refer to the test input value based on the PIT function as the realized PIT values. Rosenblatt (1952) shows that if the forecast distribution adequately reflects the distribution of the losses, the realized PIT values should form an independent and identically distributed (iid) sequence of standard uniform variables. We exploit this property to construct backtests of the realized PIT values.

Note that there are two separate groups that are interested in validating the market risk models used by the financial institutions. The first group is the financial institution themselves, where their objective is to maximize profit subject to undertaking an acceptable level of risk. The second group is the regulators, where their objective is to ensure the stability of financial market. They hope to achieve this objective by ensuring that the financial institutions set aside enough risk capital to minimize the risk of insolvency to an acceptable level. We refer to the model validation performed by the first group as internal model validation, and the second group as external model validation. The main difference between the internal and external model validation approach is due to the availability of data, since the regulators usually only have access to the data reported by the financial institutions, which is a much smaller subset of the data available to the financial institutions themselves.

A different area of model validation is the comparative backtest, where different forecast models are compared and ranked. Comparative backtest is closely linked to the concept of elicitability. When a risk measure is elicitable we can use consistent

scoring functions as the test input function to assign test input values to the risk measure estimates. Similarly, we can also rank forecast distributions using proper scoring rules as the test input function, as described in Gneiting & Raftery (2007). We refer to the test input values computed using the consistent scoring functions and proper scoring rules as the realized scores.

We can then compare different sets of risk measure estimates and forecast distributions using the realized scores, where those with lower average realized score are preferred. Note that while the realized scores are useful for model selection, we cannot perform statistical tests on these scores directly since the distribution of the realized scores depends on the distribution of the losses, which is not known.

1.3 Literature review

1.3.1 Risk measures and their properties

There is a very large literature on risk measures and their properties, and we will focus on the key references. Artzner et al. (1999) have proposes a set of desirable mathematical properties that risk measures should have. Risk measures that satisfy these properties are known as coherent risk measures. The important properties are subadditivity, which measures how much can we gain from diversifying a risky portfolio, and positive homogeneity, which requires the risk measures to scale linearly with portfolio size. Föllmer (2002) defined the larger class of convex risk measures and showed that a convex risk measure is subadditive if and only if it is positive homogeneous. See also Föllmer & Schied (2011).

Emmer et al. (2015) have provided an overview of popular risk measures, and discussed the advantages and disadvantages of these risk measures. Even though the VaR is not a coherent risk measure, due to its simplicity, it has been the dominant risk measure in banking regulation. The main issue with the use of VaR is that it is insensitive to the tail of the loss distribution. Also, while the concept of VaR is straightforward, its implementation is not. For example, when the forecast model is the HS model, there are numerous possible methods to interpolate between two data points, and different interpolation methods will lead to different VaR estimates. The

variance caused by the diversity of implementations of the same model is known as Systems Risk (Marshall & Siegel, 1997). These problems lead to a need for a better risk measure.

Acerbi & Tasche (2002) and Tasche (2002) have shown that ES is a coherent risk measure that is sensitive to the tail of the loss distribution. There have been many debates around the question of whether or not ES is amenable to direct backtesting, due to the fact that ES lacks the property of elicitability (see Section 1.3.6 for more references on the concept of elicitability). However, it is possible to backtest ES given the VaR estimates at the same level, and there are several proposed backtesting methodologies (see Section 1.3.3 for more references on this topic).

In view of the above arguments, the Basel Committee (Basel Committee on Banking Supervision, 2013, 2016) have decided to use a 10-day ES at the 97.5% level for setting trading book capital under Basel III. This decision is in line with the Swiss Solvency Test (SST), where the risk measure used is the one-year mean ES at the 99% level, while the risk measure used in solvency II remains as the 99.5% VaR. The Basel Committee have also proposed a traffic-light system to determine capital multipliers that should be applied to capital charges derived from poor risk models (see for example, Basel Committee on Banking Supervision, 2016, Appendix B). Costanzino & Curran (2016) have further extended the idea and proposed a traffic-light system analogous to the Basel system for ES.

1.3.2 Backtesting of VaR

There is a large amount of literature on traditional backtests for VaR estimates, mostly based on the binary test input function that takes the value one when a VaR exception occurs, and zero otherwise. A VaR exception refers to the event when the realized loss exceeds the VaR estimate. If the VaR at level α is consistently well estimated, the VaR exceptions should form a sequence of independent, identically distributed (iid) Bernoulli variables with success probability $1-\alpha$. See Section 3.2.1 for more details on the construction of VaR exception test using the above property. Tests that explicitly examine the serial independence of VaR exceptions are referred to as tests of conditional coverage, whereas those that do not are referred to as tests

of unconditional coverage.

Kupiec (1995) has proposed a binomial likelihood ratio test based on VaR exceptions. Christoffersen (1998) has proposed a likelihood ratio test of conditional coverage, where the alternate hypothesis is that the VaR exceptions exhibit a first-order Markov serial dependence. This test has been further studied by Davis (2013). Davé & Stahl (1998) proposed a test based on the fact that the spacings between VaR exceptions should be geometrically distributed. See also McNeil et al. (2015) for more details on the theory. Christoffersen & Pelletier (2004) refined the geometric test using the fact that a discrete geometric distribution can be approximated by a continuous exponential distribution. Engle & Manganelli (2004) have proposed a regression based test which is developed for checking the fit of the CaViaR model for dynamic quantiles. Dumitrescu et al. (2012) further develop the regression based test by considering the dynamic binary models of Kauppi & Saikkonen (2008). Berkowitz et al. (2011) have provided an overview of tests of conditional coverage. Nolde & Ziegel (2016) used the identification function (Davis, 2016) as an alternative to the binary test input function for VaR backtesting, where they refer to the tests for unconditional and conditional coverage as testing whether the VaR estimates are 'calibrated on average' and 'conditionally calibrated'.

1.3.3 Backtesting of ES

The literature on the backtesting of ES is much smaller. In most existing literature, the joint identification function for VaR and ES (Nolde & Ziegel, 2016) is used as the test input function. McNeil & Frey (2000) suggest a bootstrap hypothesis test on the test inputs, which they refer to as the violation residuals. These measure the discrepancy between the realized losses and the expected shortfall estimates conditional on VaR exception occurring, which should form a sample from a distribution with mean zero. Acerbi & Szekely (2014) look at similar statistics and suggest constructing a test using Monte Carlo methods. Nolde & Ziegel (2016) proposed to test the violation residuals by constructing a Z-test statistic. Recently Costanzino & Curran (2015) have proposed a Z-test for a discretized version of expected shortfall. See Clift et al. (2015) for an empirical comparison of the above backtests.

1.3.4 Backtesting of realized PIT values

Diebold et al. (1998) have shown how realized PIT values can be used to evaluate the overall quality of forecast distributions. In Diebold et al. (1999), the authors extended the evaluation of forecast distributions to the multivariate case. Blum (2004) proposed a method based on realized PIT values to deal with problems relating to overlapping forecast intervals and multiple forecast horizons. Berkowitz (2001) proposed a test on the tail of the forecast distribution based on the idea of truncating realized PIT values above a level α. Other relevant literature include Kerkhof & Melenberg (2004) who proposed to backtest VaR and ES by applying a functional delta method to the empirical distribution function of the realized PIT and Zumbach (2006) who refers to realized PIT values as probtiles.

1.3.5 Empirical results

There is some literature that contains empirical backtesting results of banks. O'Brien & Szerszen (2014) found that a large number of banks do not pass the conditional coverage test for VaR estimates. Pérignon & Smith (2010) reported that 73% of US and international banks use the historical simulation (HS) method as their forecast model. The HS method is a non-parametric method which estimates the conditional loss distribution by re-sampling historical risk-factor changes. Since the dynamics of the loss distribution are ignored, this method may lead to dependencies in the series of VaR exceptions. Some banks reported the use of filtered historical simulation (FHS), which is the dynamic version of HS that adjusts for volatility. For more detail on FHS, see for example, Barone-Adesi et al. (1998), Hull & White (1998) and McNeil et al. (2015). Pritsker (2006) has criticized the use of the HS and FHS methodology. We will pay particular attention to these methods due to their popularity in the industry.

1.3.6 Elicitability theory

The concept of elicitability originates from the work by Savage (1971), further developed by Osband & Reichelstein (1985). The name was coined by Lambert et al.

(2008) and popularized by Gneiting (2011). Gneiting (2011) showed that ES is not an elicitable risk measure, whereas VaR is. See also, for example, Gneiting (2011); Ziegel (2013); Bellini & Bignozzi (2015). However, ES satisfies more general notions of elicitability, such as conditional elicitability (Emmer et al., 2015) and joint elicitability (Fissler & Ziegel, 2015). Acerbi & Székely (2016) have introduced a new concept of "backtestability", which is satisfied in particular by expected shortfall. There is a very large literature on comparative forecasting performance based on elicitability theory, which include Diebold & Mariano (1995), West (1996) and Giacomini & White (2006).

1.4 Structure of the thesis

In this thesis, we will summarize the general concepts of hypothesis testing in Chapter 2. In Chapter 3, we will summarize the existing tests for VaR and realized PIT values, as well as introducing some new tests. In particular, we develop a framework for testing weighted transformations of realized PIT values, and show that many of the existing VaR and realized PIT backtests fit into this framework. We then perform simulation studies in Chapter 4 to better understand the size and power of the tests described in Chapter 3. In Chapter 5, we extend the general framework introduced in Chapter 3 to explicitly tests for serial independence of the realized PIT values. The tests described in this chapter may be more useful in practice since many of the banks failed to model the dynamics of the losses appropriately (O'Brien & Szerszen, 2014). The simulation results for the tests described in Chapter 5 are shown in Chapter 6. Finally, some existing and new ideas based on elicitability theory are explored in Chapter 7.

Chapter 2 A review of hypothesis testing

In this chapter, we summarize some of the general concepts of hypothesis testing that we will be using frequently. We will follow the treatment given in Casella & Berger (2001) closely.

Let $X_1, \ldots, X_n$, be an iid random sample with probability density function (pdf) or probability mass function (pmf) $f(x|\theta)$. Fow now, we assume θ to be a scalar with parameter space Θ. We are interested in testing whether

$$H_0 : \ \theta \in \Theta_0 \ \text{ vs. } \ H_1 : \theta \in \Theta_0^c \,, \tag{2.1}$$

where Θ_0 is some subset of the parameter space and Θ_0^c is its complement. H_0 and H_1 are known as the null and alternative hypothesis.

To determine whether we accept H_0 (and thus reject H_1), we require a test statistic $W(\boldsymbol{X}) = W(X_1, \ldots, X_n)$, which is a function of the sample. We reject H_0 when $W(\boldsymbol{X}) \in R$, where R is known as the rejection region. The performance of a hypothesis test depends on the choice of test statistic $W(\boldsymbol{X})$, which we evaluate based on two types of error:

- Type I error: Probability of rejecting H_0 when it is true.
- Type II error: Probability of accepting H_0 when it is false.

Type I and type II errors are summarized in the power function, defined by

$$\beta(\theta) = P(W(\boldsymbol{X}) \in R|\theta) \,. \tag{2.2}$$

We say that a test with power function $\beta(\theta)$ has size κ if $\sup_{\theta \in \Theta_0} \beta(\theta) = \kappa$, for $0 \leqslant \kappa \leqslant 1$.

Suppose we denote class C to be the class of all size κ test. A test in class C should be preferred if it has a small Type II error, or large power function for $\theta \in \Theta_0^c$.

Sometimes it may be useful to report the p-value of the test statistic. Let $W(\boldsymbol{X})$ be a test statistic such that we reject large values of $W(\boldsymbol{X})$. We follow Casella &

Berger (2001) and define the test p-value for a realized sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ by

$$p(\boldsymbol{x}) = \sup_{\theta \in \Theta_0} P_\theta\big(W(\boldsymbol{X}) \geqslant W(\boldsymbol{x})\big)\,, \tag{2.3}$$

which is the maximum probability of observing a result as extreme as $W(\boldsymbol{x})$ under H_0.

In many cases, the distribution of the test statistic $W(\boldsymbol{X})$ is approximated using asymptotic theory. We denote by $\tilde{\beta}(\theta)$ the derived power function using (2.2) by assuming large n. In such cases, we say that the test has approximately size κ if $\sup_{\theta \in \Theta_0} \tilde{\beta}(\theta) = \kappa$. Generally, when n is small, $\sup_{\theta \in \Theta_0} \beta(\theta) \neq \kappa$. We loosely refer to this discrepancy as the size performance of the test.

2.1 Some useful statistical concepts

In this section, we state some well known statistical concepts that will be useful in the following sections.

We define the log-likelihood function $l(\theta|x) = \log f(x|\theta)$ and the score function $u(\theta|x) = \frac{\partial}{\partial \theta} l(\theta|x)$. The log-likelihood for a given realization $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ is $L_n(\theta|\boldsymbol{x}) = \frac{1}{n}\sum_{t=1}^{n} l(\theta|x_t)$. We recall the following fundamental results in statistics:

Strong Law of Large Numbers (SLLN) Assuming that $\mathrm{E}\,(X) < \infty$, the sample average $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges almost surely to the expectation $\mathrm{E}\,(X)$.

Central Limit Theorem (CLT) Assuming that $\mathrm{var}(X) < \infty$, $\sqrt{\frac{n}{\mathrm{var}(X)}}(\bar{X}_n - \mathrm{E}\,(X))$ converges in distribution to a standard normal random variable. A Z-test is a test whose test statistic is constructed based on the CLT result.

Fisher information Under suitable regularity conditions (we refer to Casella & Berger (2001) p. 516), we have that $\mathrm{E}_\theta(u(\theta|X)) = 0$ and $\mathrm{var}_\theta(u(\theta|X)) = I(\theta)$, where $I(\theta)$ is known as the Fisher information. Under these regularity conditions, it can be shown that

$$I(\theta) = -\mathrm{E}_\theta\left(\frac{\partial}{\partial \theta} u(\theta|X)\right)\,. \tag{2.4}$$

We will assume that these regularity conditions apply throughout the rest of

the thesis. The Fisher information is used in the score test and Wald test, which will be described shortly.

2.1.1 Score test

We denote by $L'_n(\theta|\boldsymbol{x})$ and $L''_n(\theta|\boldsymbol{x})$ the first and second partial derivatives of the log-likelihood with respect to θ, with $L'_n(\theta|\boldsymbol{x}) = \frac{1}{n}\sum_{t=1}^{n} u(\theta|x_t)$. Direct application of the CLT leads to

$$Z_{\text{score}} = \sqrt{\frac{n}{I(\theta)}} L'_n(\theta|\boldsymbol{x})\,, \tag{2.5}$$

where Z_{score} is known as the score test statistic, and is asymptotically standard normal.

2.1.2 Wald test

Using Taylor's approximation,

$$L'_n(\hat{\theta}|\boldsymbol{x}) \approx L'_n(\theta|\boldsymbol{x}) + L''_n(\theta|\boldsymbol{x})(\hat{\theta} - \theta)\,. \tag{2.6}$$

Since $L'_n(\hat{\theta}|\boldsymbol{x}) = 0$, and $L''_n(\theta|\boldsymbol{x})$ converges to $-I(\theta)$ by SLLN due to (2.4), $L'_n(\theta|\boldsymbol{x}) \approx I(\theta)(\hat{\theta} - \theta)$ for large n. Plugging in this result into (2.5), and we obtain the Wald test statistic

$$Z_{\text{Wald}} = \sqrt{nI(\theta)}(\hat{\theta} - \theta)\,, \tag{2.7}$$

which is asymptotically standard normal. Note that there are several definitions for the Wald test. Here, we define the Wald test to be the test based on application of the CLT to the maximum likelihood estimator $\hat{\theta}$.

It is quite common to replace the Fisher information $I(\theta)$ in (2.7) with the observed Fisher information, since in many cases it will improve the size performance of the test. For example, two variants of observed Fisher information are $-L''_n(\hat{\theta}|\boldsymbol{x})$ and $\frac{1}{n}\sum_{t=1}^{n}(\frac{\partial}{\partial\theta}\log f(x_t|\hat{\theta}))^2$, where we denote by $\hat{\theta}$ the (unconstrained) maximum likelihood estimator for $L_n(\theta|\boldsymbol{x})$.

2.1.3 Likelihood ratio test

The likelihood ratio test (LRT) statistic is given by

$$S_{\text{LRT}} = -2\sum_{i=1}^{n}(l(\hat{\theta}_0|\boldsymbol{x}) - l(\hat{\theta}|\boldsymbol{x}))\,, \tag{2.8}$$

where $\hat{\theta}_0 = \arg\max_{\theta\in\Theta_0} L_n(\theta|\boldsymbol{x})$ is the constrained maximum likelihood estimator.

2.2 Two-sided hypothesis

In this section we will describe how to use the score test, Wald test and LRT to test the two-sided hypothesis $H_0 :\ \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. We will also give the generalization to a test of multiple parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ with the two-sided hypothesis $H_0 :\ \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

2.2.1 Score test

Under H_0, using (2.5) with $\theta = \theta_0$, we have that for large enough n the score test statistic

$$Z_{\text{score}} = \sqrt{\frac{n}{I(\theta_0)}}L_n'(\theta_0|\boldsymbol{x}) \sim N(0,1)\,. \tag{2.9}$$

To test for the two-sided hypothesis, we use the fact that Z^2_{score} converges to a χ^2_1 distribution for large n, and reject the null hypothesis when $Z^2_{\text{score}} > c$. To obtain a test that has approximately size κ, we set $c = F^{-1}_{\chi^2_1}(1-\kappa)$, where $F^{-1}_{\chi^2_k}$ denotes the inverse of a chi-square distribution with k degree of freedom.

We can easily extend the above result to when $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is k-dimensional. For large enough n the two-sided score test statistic

$$S_{\text{score}} = n(L_n'(\boldsymbol{\theta}_0|\boldsymbol{x}))^T I(\boldsymbol{\theta}_0)^{-1} L_n'(\boldsymbol{\theta}_0|\boldsymbol{x}) \sim \chi^2_k\,, \tag{2.10}$$

and to obtain a test that has approximately size κ, we reject the null hypothesis when $S_{\text{score}} > F^{-1}_{\chi^2_k}(1-\kappa)$.

2.2.2 Wald test

Under H_0, using (2.7) with $\theta = \theta_0$, we have that for large enough n the Wald test statistic

$$Z_{\text{Wald}} = \sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \sim N(0, 1) \,. \tag{2.11}$$

To obtain a test that has approximately size κ, we reject the null hypothesis when $Z^2_{\text{Wald}} > F^{-1}_{\chi^2_1}(1 - \kappa)$.

When $\boldsymbol{\theta}$ is k-dimensional, for large enough n the two-sided Wald test statistic

$$S_{\text{Wald}} = n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T I(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \sim \chi^2_k \,. \tag{2.12}$$

To obtain a test that has approximately size κ, we reject the null hypothesis when $S_{\text{Wald}} > F^{-1}_{\chi^2_k}(1 - \kappa)$.

2.2.3 Likelihood ratio test

By applying a Taylor's approximation to (2.8), we obtain

$$S_{\text{LRT}} \approx -2n(\hat{\theta}_0 - \hat{\theta})^T L'_n(\hat{\theta}|\boldsymbol{x}) - n(\hat{\theta}_0 - \hat{\theta})^T L''_n(\hat{\theta}|\boldsymbol{x})(\hat{\theta}_0 - \hat{\theta}) \,. \tag{2.13}$$

Under H_0, we have that $\hat{\theta}_0 = \theta_0$ and $L''_n(\hat{\theta}|\boldsymbol{x})$ converges to $-I(\theta_0)$ by SLLN and (2.4). Using the fact that $L'_n(\hat{\theta}|\boldsymbol{x}) = 0$, and the two-sided Wald test statistic result in (2.12), the RHS of (2.8) converges to a χ^2_1 random variable for large n, and to obtain a test that has approximately size κ, we reject the null hypothesis when $S_{\text{LRT}} > F^{-1}_{\chi^2_1}(1 - \kappa)$.

We can easily extend the above result to when $\boldsymbol{\theta}$ is k-dimensional, in which case under H_0 for large enough n, the LRT statistic

$$S_{\text{LRT}} = -2 \sum_{i=1}^{n} (l(\boldsymbol{\theta}_0|\boldsymbol{x}) - l(\hat{\boldsymbol{\theta}}|\boldsymbol{x})) \sim \chi^2_k \,. \tag{2.14}$$

To obtain a test that has approximately size κ, we reject the null hypothesis when $S_{\text{LRT}} > F^{-1}_{\chi^2_k}(1 - \kappa)$.

2.3 One-sided hypothesis

In this section we will describe how to construct a one-sided hypothesis $H_0 : \theta \leqslant \theta_0$ vs. $H_1 : \theta > \theta_0$. Devising a one-sided test is more complicated than the two-sided case, and we will only consider the Wald test and LRT in the case when θ is one-dimensional.

2.3.1 Wald test

From (2.7), we know that for large n, $\hat{\theta} \sim N\left(\theta, \frac{1}{nI(\theta)}\right)$. Following the arguments in Casella & Berger (2001) p. 398, the supremum in (2.3) always occur at θ_0. Hence, we only need to consider the case $\theta = \theta_0$. Using (2.7) with $\theta = \theta_0$, we have that for large enough n the Wald test statistic

$$Z_{\text{Wald}} = \sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \sim N(0, 1) \,, \tag{2.15}$$

and we reject H_0 when $Z_{\text{Wald}} > \Phi^{-1}(1 - \kappa)$ to obtain a test that has approximately size κ.

2.3.2 Likelihood ratio test

Testing the one sided hypothesis for LRT is more tricky, and we will only consider the case when θ is one-dimensional. Suppose we reject S_{LRT} in (2.8) when $S_{\text{LRT}} > c$ for some c. Recall Taylor's approximation

$$S_{\text{LRT}} \approx -2n(\hat{\theta}_0 - \hat{\theta})^T L_n'(\hat{\theta}|\boldsymbol{x}) - n(\hat{\theta}_0 - \hat{\theta})^T L_n''(\hat{\theta}|\boldsymbol{x})(\hat{\theta}_0 - \hat{\theta}) \,. \tag{2.16}$$

If $\theta < \theta_0$, then for large enough n we have that $\hat{\theta}_0 = \hat{\theta}$, which leads to $S_{\text{LRT}} = 0$. If $\theta > \theta_0$, as n tend to infinity, S_{LRT} will tend to infinity as well. To obtain a test with approximate size κ, we need to find c such that

$$\lim_{n\to\infty} P(S_{\text{LRT}} > c|\theta) = \begin{cases} 0 & \theta < \theta_0 \,, \\ \kappa & \theta = \theta_0 \,, \\ 1 & \theta > \theta_0 \,. \end{cases} \tag{2.17}$$

To determine c, we only need to consider the case when $\theta = \theta_0$, i.e. the equation

$$P(\hat{\theta} < \theta_0)P(S_{\text{LRT}} > c|\hat{\theta} < \theta_0,\ \theta = \theta_0) + P(\hat{\theta} \geqslant \theta_0)P(S_{\text{LRT}} > c|\hat{\theta} \geqslant \theta_0,\ \theta = \theta_0) = \kappa\,. \tag{2.18}$$

When $\theta = \theta_0$, from (2.7), we know that $\hat{\theta}$ is symmetrically distributed around θ_0, and hence $P(\hat{\theta} < \theta_0) = P(\hat{\theta} \geqslant \theta_0) = 0.5$. Given $\hat{\theta} < \theta_0$, we have that $\hat{\theta}_0 = \hat{\theta}$ which leads to $S_{\text{LRT}} = 0$ and $P(S_{\text{LRT}} > c|\hat{\theta} < \theta_0,\ \theta = \theta_0) = 0$.

In the case when $\hat{\theta} \geqslant \theta_0$, we will assume that the likelihood in the region Θ_0 is increasing so that $\hat{\theta}_0 = \theta_0$, in which case S_{LRT} is the two-sided LRT statistic, and using the results from Section 2.2.3, for large enough n, we have that $S_{\text{LRT}} \sim \chi^2_1$ so that $P(S_{\text{LRT}} > c|\hat{\theta} \geqslant \theta_0,\ \theta = \theta_0) = 1 - F_{\chi^2_1}(c)$. Hence, we have that $c = F^{-1}_{\chi^2_1}(1-2\kappa)$. For more details, we refer to Chen & Shi (2011).

Chapter 3 Backtesting VaR, ES and realized PIT values

To construct backtests for a set of Value-at-Risk (VaR) or expected shortfall (ES) estimates, we compare the estimates with the realized losses, observed ex post, and assign values to the estimates, which we refer to as test inputs values, based on some function, which we refer to as the test input function. We then perform hypothesis tests on the test inputs using techniques described in the previous chapter.

Similarly, we can construct backtests for a forecast distribution by assigning test inputs values to the realized losses computed using the forecast distribution. For the evaluation of forecast distributions, we will focus on the case when the test input function is the probability-integral-transform (PIT), in which case we refer to the test input values as the realized PIT values.

In this chapter, we will focus on static tests, and later in Chapter 5, we will describe some dynamic tests. We refer to tests which explicitly test for the serial independence of the test inputs as dynamic tests (or conditional coverage tests), and those which do not as static tests (or unconditional coverage tests).

3.1 VaR, spectral risk measures and ES

We will summarize the VaR and ES risk measures. We denote by $\mathcal{F}_t$ the filtration at time t, which represents the information available to the modeller at time t, and we denote the conditional loss distribution given information up to time $t-1$ by

$$F_t(x) = \mathrm{P}\left(L_t \leqslant x \mid \mathcal{F}_{t-1}\right), \tag{3.1}$$

which we will assume in the sequel to be a continuous distribution for all t.

The VaR at level α for F_t is defined as the generalized inverse of F_t at α, given by

$$\mathrm{VaR}_{\alpha,t} = F_t^{\leftarrow}(\alpha) = \inf\{x \in \mathbb{R} : F_t(x) > \alpha\}. \tag{3.2}$$

When F_t is continuous, the VaR is simply the ordinary inverse of F_t.

Spectral risk measures are weighted integrals of VaR at multiple levels, where the

weight function g is required to satisfy certain constraints. Following Costanzino & Curran (2015), we say that g is an admissible risk spectrum if

i g is non-negative,

ii g is non-decreasing,

iii $\int_0^1 g(u)\mathrm{d}u = 1\,.$

A spectral risk measure $\mathcal{M}_{g,t}$ with an admissible risk spectrum g is defined as

$$\mathcal{M}_{g,t} = \int_0^1 g(u)\,\mathrm{VaR}_{u,t}\,\mathrm{d}u\,. \tag{3.3}$$

The ES at level α for F_t, denoted by $\mathrm{ES}_{\alpha,t}$, is given by

$$\mathrm{ES}_{\alpha,t} = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_{u,t}\,\mathrm{d}u\,, \tag{3.4}$$

which is a special case of the spectral risk measure, with $g(u) = \frac{1}{1-\alpha} I_{\{\alpha \leqslant u \leqslant 1\}}$.

Spectral risk measures have become a popular area of research as Acerbi (2002) has shown that spectral risk measures are always coherent. See Artzner et al. (1999) for more details on the properties of coherent risk measures and why they are desirable. Another advantage of spectral risk measures is that they can be related to the concept of risk aversion, due to the use of non-decreasing weight functions. It is interesting to note that Heyde et al. (2007) have proposed a new risk measure, which they refer to as the natural risk statistic, which is characterized by a new set of axioms. The natural risk statistic includes the tail conditional median which is shown to be more robust than ES. Also, the natural risk statistic includes VaR as a special case.

3.2 Tests for VaR exceptions

Let $\mathcal{G}_t \subset \mathcal{F}_t$ be the sigma algebra based on finite history used to compute the forecast distribution $\widehat{F}_t$. Ideally, we would prefer to test the null hypothesis

$$\mathrm{E}\left(d\big(\widehat{\mathrm{VaR}}_{\alpha,t}, F_t^{-1}(\alpha)\big) \mid \mathcal{G}_{t-1}\right) = 0\,, \tag{3.5}$$

where d is some distance measure, and $\widehat{\mathrm{VaR}}_{\alpha,t}$ is the VaR estimate at level α computed using the forecast distribution $\widehat{F}_t$. In the tests that we study subsequently,

we will adopt the same philosophy as Giacomini & White (2006), where we wish to evaluate the forecasting method, i.e. evaluate in addition to the forecast model, the estimation procedure and the data used for estimation. In other words, a poorly chosen information set $\mathcal{G}_{t-1}$ should be penalized.

If we wish to evaluate the forecast model only, we will need to take into account model estimation error. Such a correction will be complex, as it needs to take into account a number of factors, including the ratio of the estimation window to the forecast horizon, model estimation method and the choice of test input function. See for example, West (1996). Hence, such a correction is not feasible for external model validation. For internal model validation, estimation error can be important, and the modeller may wish to allow for it when constructing backtests.

An important point made in West (1996) is that the size of correction depends on the ratio of the estimation window to the forecast horizon. For our simulation studies later, since the forecast horizon is one day, and the forecast models are estimated using a rolling window procedure, this ratio is small when the estimation window is large enough, which means that the effect of the parameter estimation error on the tests is rather small.

We will refer to tests in the case when the distance measure in (3.5) is

$$d(x, y) = x - y \tag{3.6}$$

as bias-based tests, since under the null hypothesis $\widehat{\mathrm{VaR}}_{\alpha,t}$ is an unbiased estimator of $F_t^{-1}(\alpha)$. Such a test is difficult to carry out, as it requires the knowledge of the true loss distribution F_t, which is usually not known.

Hence, in practice, we will compromise by using a exception-based test, where we refer to the event $\{L_t > \mathrm{VaR}_{\alpha,t}\}$ as a VaR exception at level α, and we compare α with $I_{\{L_t \leqslant \widehat{\mathrm{VaR}}_{\alpha,t}\}}$ to test the null hypothesis

$$\mathrm{E}\left(\alpha - I_{\{L_t \leqslant \widehat{\mathrm{VaR}}_{\alpha,t}\}} \mid \mathcal{G}_{t-1}\right) = 0\,. \tag{3.7}$$

In the sequel, for notation simplicity, we will drop the conditioning on $\mathcal{G}_{t-1}$. When we construct a test, we approximate the expectation with the observed exception rate. When the amount of data used to compute the observed exception rate is small,

the observed exception rate will likely be a poor approximation to the expectation, hence tests which are based on asymptotic theory, such as the LRT in Section 2.2.3, will have poor size performance. Size correction techniques can be performed in such cases to improve size performance. In the case of LRT, we can perform, for example, Bartlett-type corrections (Bartlett, 1937).

For $t \in \mathbb{N}$ we define $(I_{\alpha,t})$ to be the exception process at level α, where $I_{\alpha,t} = I_{\{L_t > \mathrm{VaR}_{\alpha,t}\}}$. Christoffersen (1998) show that the sequence $(I_{\alpha,t})$ should satisfy the unconditional coverage hypothesis, which states that $\mathrm{E}\,(I_{\alpha,t}) = 1 - \alpha$ for all t, as well as the conditional coverage hypothesis, which states that $I_{\alpha,t}$ is independent of $I_{\alpha,s}$ for $s \neq t$.

We denote by $S_\alpha(v, l)$ the test input function for VaR at level α. We will use the binary test input function $S_\alpha(v, l) = I_{\{l>v\}}$, and for a given sequence of VaR estimates $(\widehat{\mathrm{VaR}}_{\alpha,t})$ with corresponding realized losses (L_t), the test input sequence is $\big(S_\alpha(\widehat{\mathrm{VaR}}_{\alpha,t}, L_t)\big)$. There are other choices of test input functions. For example, Nolde & Ziegel (2016) have proposed to use the identification function for VaR at level α, denoted by $h_\alpha(v, l) = I_{\{l>v\}} - (1 - \alpha)$, as the test input function. See for example, Davis (2013) for more details on the construction and theory of identification functions. The test input sequence $\big(h_\alpha(\widehat{\mathrm{VaR}}_{\alpha,t}, L_t)\big)$ is said to satisfy the unconditional calibration hypothesis if $\mathrm{E}\,\big(h_\alpha(\widehat{\mathrm{VaR}}_{\alpha,t}, L_t)\big) = 0$ for all t. The test input sequence is said to satisfy the conditional calibration hypothesis if $\mathrm{E}\,\big(h_\alpha(\widehat{\mathrm{VaR}}_{\alpha,t}, L_t) \mid \mathcal{F}_{t-1}\big) = 0$ for all t.

We denote by $\mathrm{VaR}_{\boldsymbol{\alpha},t} = (\mathrm{VaR}_{\alpha_1,t}, \ldots, \mathrm{VaR}_{\alpha_N,t})$ a series of VaR at ordered levels $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$. We set $\alpha_0 = 0$ and $\alpha_{N+1} = 1$, and we define the exception indicator at the level α_i and at time t by $I_{\alpha_i,t} = I_{\{L_t > \mathrm{VaR}_{\alpha_i,t}\}}$. We can simultaneously test the VaR estimates at N levels based on the exception process. We define

$$C_t = \sum_{i=1}^{N} I_{\alpha_i,t}\,, \tag{3.8}$$

where the sequence (C_t) counts the number of VaR levels that are breached, and should satisfy the unconditional coverage hypothesis, which states that $\mathrm{P}\,(C_t \leqslant i) = \alpha_{i+1},\ i = 0, \ldots, N$ for all t, and the conditional coverage hypothesis, which states that C_t is independent of C_s for $s \neq t$.

The unconditional coverage property can also be written as

$$C_t \sim \text{MN}(1, (\alpha_1 - \alpha_0, \ldots, \alpha_{N+1} - \alpha_N)) \tag{3.9}$$

for all t, where $\text{MN}(n, (p_0, \ldots, p_N))$ denotes the multinomial distribution with n trials, each of which may result in one of $N+1$ outcomes $\{0, 1, \ldots, N\}$ with corresponding probabilities $p_0, \ldots, p_N$ which satisfy $\sum_{i=0}^{N} p_i = 1$.

3.2.1 Binomial tests

We first consider the case when $N = 1$, with $\alpha_1 = \alpha$. Under the unconditional coverage hypothesis and conditional coverage hypothesis, the series $(C_1, \ldots, C_n)$ are iid Bernoulli variables with success probability $p_0 = 1 - \alpha$. We consider two-sided tests with hypothesis

$$H_0 : \quad p = p_0 \quad \text{vs.} \quad H_1 : \quad p \neq p_0 \,, \tag{3.10}$$

and one-sided tests with hypothesis

$$H_0 : \quad p \leqslant p_0 \quad \text{vs.} \quad H_1 : \quad p > p_0 \,. \tag{3.11}$$

We define the observed exception rate to be $\hat{p} = \frac{O_1}{n}$, where $O_1 = \sum_{t=1}^{n} I_{\{C_t=1\}}$. The binomial score test statistic is given by

$$Z_{\text{score}} = \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)}} \,, \tag{3.12}$$

which is asymptotically standard normal distributed. Note that for the binomial test, the Wald test statistic is the same as the score test statistic.

The binomial LRT statistic is given by

$$S_{\text{LRT}} = -2 \ln \left(\frac{(1-\hat{p}_0)^{n-O_1} (\hat{p}_0)^{O_1}}{(1-\hat{p})^{n-O_1} (\hat{p})^{O_1}} \right) . \tag{3.13}$$

See for example, Kupiec (1995). To construct a test with approximate size κ, in the two-sided case, we set $\hat{p}_0 = p_0$, and the null is rejected if $S_{\text{LRT}} > F^{-1}_{\chi^2_1}(1-\kappa)$. For the one-sided case, we set $\hat{p}_0 = \hat{p}$ if $\hat{p} \leqslant p_0$, and $\hat{p}_0 = p_0$ if $\hat{p} > p_0$, and the null is rejected if $S_{\text{LRT}} > F^{-1}_{\chi^2_1}(1-2\kappa)$. See Section 2.3.2 for more details on the one-sided LRT.

3.2.2 Multinomial tests

We now consider the case when the number of VaR levels $N \geqslant 2$. For a series $(C_1, \ldots, C_n)$, we define the observed cell count to be $O_i = \sum_{t=1}^{n} I_{\{C_t = i\}}$, for $i = 0, 1 \ldots, N$. Under the unconditional coverage hypothesis and conditional coverage hypothesis, $(O_0, \ldots, O_N)$ should follow a multinomial distribution

$$(O_0, \ldots, O_N) \sim \mathrm{MN}(n, (\alpha_1 - \alpha_0, \ldots, \alpha_{N+1} - \alpha_N)) \,. \tag{3.14}$$

Suppose we define $0 = \theta_0 < \theta_1 < \cdots < \theta_N < \theta_{N+1} = 1$ to be an arbitrary sequence of parameters, and we consider the model

$$(O_0, \ldots, O_N) \sim \mathrm{MN}(n, (\theta_1 - \theta_0, \ldots, \theta_{N+1} - \theta_N)) \,. \tag{3.15}$$

We then construct the hypotheses

$$\begin{aligned} H_0 : \quad & \theta_i \;\; = \alpha_i \quad \text{for } i = 1, \ldots, N \,, \\ H_1 : \quad & \theta_i \;\; \neq \alpha_i \quad \text{for at least one } i \in \{1, \ldots, N\} \,. \end{aligned} \tag{3.16}$$

Cai & Krishnamoorthy (2006) studied five different tests to test for the hypothesis in (3.16) We will consider three of them:

Pearson chi-squared test of (Pearson, 1900). The test statistic is given by

$$S_N = \sum_{i=0}^{N} \frac{(O_i - n(\alpha_{i+1} - \alpha_i))^2}{n(\alpha_{i+1} - \alpha_i)} \,, \tag{3.17}$$

where under the H_0 in (3.16), S_N is asymptotically χ^2_N distributed. It is well known that the accuracy of this test increases as the minimum interval size $\min\limits_{0 \leqslant i \leqslant N} n(\alpha_{i+1} - \alpha_i)$ increases, and decreases as the number of levels N increases. Note that when $N = 1$, the Pearson chi-squared test statistic is the same as the two-sided binomial score test statistic Z^2_{score}, where Z_{score} is defined in (3.12).

Nass test. Nass (1959) studied an improved approximation to the distribution of the statistic S_N defined in (3.17), first introduced by Vessereau (1958), who proposed to find c and v such that the first two moments of $c\, S_N$ matches the first two moments of the χ^2_v random variable, i.e.

$$c\, S_N \underset{H_0}{\overset{d}{\sim}} \chi^2_\nu \,, \quad \text{with} \quad c = \frac{2\,\mathrm{E}\,(S_N)}{\mathrm{var}(S_N)} \text{ and } \nu = c\,\mathrm{E}\,(S_N) \,. \tag{3.18}$$

Pearson (1932) show that

$$\mathrm{E}\,(S_N) = N\,, \quad \text{and} \quad \mathrm{var}(S_N) = 2N - \frac{N^2+4N+1}{n} + \frac{1}{n}\sum_{i=0}^{N}\frac{1}{\alpha_{i+1}-\alpha_i}\,. \tag{3.19}$$

We will refer to the size-corrected chi-squared test as the Nass test.

Discrete probitnormal LRT. The multinomial LRT test statistic is given by

$$\tilde{S}_N = 2\sum_{i=0}^{N} O_i \ln\left(\frac{\hat{\theta}_{i+1}-\hat{\theta}_i}{\alpha_{i+1}-\alpha_i}\right)\,. \tag{3.20}$$

Under the multinomial model

$$(O_0,\ldots,O_N) \sim \mathrm{MN}\left(n, (\theta_1-\theta_0,\ldots,\theta_{N+1}-\theta_N)\right)\,, \tag{3.21}$$

the maximum likelihood estimator of the multinomial probabilities are given by $\hat{\theta}_{i+1}-\hat{\theta}_i = O_i/n$. When cell probabilities are small, we may obtain $O_i = 0$, which leads to an undefined test statistic. We propose a different version of the LRT to the one described in Cai & Krishnamoorthy (2006), which we will refer to as the discrete probitnormal LRT. We consider the multinomial model where the parameters are given by

$$\theta_i = \Phi\left(\frac{\Phi^{-1}(\alpha_i)-\mu}{\sigma}\right), \quad i = 1,\ldots,N\,, \tag{3.22}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$ and Φ denotes the standard normal distribution function. In this case, the parameter estimates are

$$\hat{\theta}_{i+1}-\hat{\theta}_i = \Phi\left(\frac{\Phi^{-1}(\alpha_{i+1})-\hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{\Phi^{-1}(\alpha_i)-\hat{\mu}}{\hat{\sigma}}\right)\,, \tag{3.23}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the MLEs under H_1. For this model, the problem of zero estimated cell probabilities does not arise.

We test the null hypothesis

$$H_0: \quad \mu = 0 \;\text{ and }\; \sigma = 1 \quad \text{vs.} \quad H_1: \quad \mu \neq 0 \;\text{ or }\; \sigma \neq 1\,, \tag{3.24}$$

and the test statistic $\tilde{S}_N$ is asymptotically χ^2_2 distributed.

3.3 Test for expected shortfall

In most existing literature, backtests for expected shortfall is based on the joint identification function for VaR and ES, which we denote as

$$h_\alpha(e, v, l) = I_{\{l>v\}} \frac{l - v}{1 - \alpha} - (e - v)\,. \tag{3.25}$$

For a given sequence of VaR estimates $(\widehat{\mathrm{VaR}}_{\alpha,t})$ and ES estimates $(\widehat{\mathrm{ES}}_{\alpha,t})$, we define the corresponding test input sequence to be (s_t), where $s_t = h_\alpha(\widehat{\mathrm{ES}}_{\alpha,t}, \widehat{\mathrm{VaR}}_{\alpha,t}, L_t)$. McNeil & Frey (2000) refer to (s_t) as the violation residuals, which measure the discrepancy between the realized losses and the expected shortfall estimates on days when VaR exception occurs, and should form a sample from a distribution with mean zero.

The unconditional calibration hypothesis is

$$H_0: \quad \mathrm{E}\,(s_t) = 0 \quad \text{vs.} \quad H_1: \quad \mathrm{E}\,(s_t) \neq 0\,, \tag{3.26}$$

and for $t = 1, \ldots, n$, we compute the Z-test statistic

$$Z_{\mathrm{ES}} = \frac{\sum_{t=1}^{n} s_t}{\sqrt{\sum_{t=1}^{n} s_t^2}}\,, \tag{3.27}$$

and reject H_0 when $Z_{\mathrm{ES}}^2 > F_{\chi_1^2}^{-1}(1 - \kappa)$ to obtain a test with approximate size κ. See for example, Nolde & Ziegel (2016) for more details on the construction of this test.

3.4 Realized PIT values

We define the process (U_t) by $U_t = F_t(L_t)$ using the probability integral transform (PIT). Under the assumption that F_t is continuous for all t, Rosenblatt (1952) showed that (U_t) should form a series of iid standard uniform variables.

Realized PIT values are the corresponding variables (P_t) obtained by setting $P_t = \widehat{F}_t(L_t)$. They are estimates of the probability of observing a realized loss no more extreme than L_t using the forecast model $\widehat{F}_t$. Assuming that $\widehat{F}_t$ is well calibrated and close to F_t for all t, we would expect the realized PIT values to behave like an iid sample of standard uniform variables.

Realized PIT values contain information about VaR exceptions at any level α. To see this, we note that

$$P_t \geqslant \alpha \iff L_t \geqslant \widehat{\mathrm{VaR}}_{\alpha,t}\,. \tag{3.28}$$

(3.28) always holds for any forecast model $\widehat{F}_t$. The weak inequalities can be replaced by strict inequalities if $\widehat{F}_t$ is strictly increasing and continuous. In the sequel we will define the PIT-based VaR exception at level α to be the event $\{P_t > \alpha\}$.

Since realized PIT values contains more information than the VaR exception at a single level α, we would expect well-designed tests based on realized PIT values to be more powerful in detecting deficiencies in the forecast models $\widehat{F}_t$.

3.4.1 A unified framework for tests based on realized PIT values

In the sequel we will focus on tests based on the transformations of realized PIT values given by

$$W_{v,t} = \int_0^1 I_{\{P_t > u\}} \mathrm{d}v(u)\,, \tag{3.29}$$

where v is a measure defined on the interval $[0,1]$ whose role is to apply weight to the PIT-based VaR exception at different levels.

We are interested in testing whether the behavior of $(W_{v,t})$ is consistent with the behavior that would be expected if (P_t) did indeed form an iid sample from a standard uniform distribution. We denote by F_W the distribution function of $(W_{v,t})$ in (3.29) when P_t is uniform. We are interested in testing the null hypothesis

$$H_0: \quad (W_{v,t}) \text{ is an iid series with distribution function } F_W. \tag{3.30}$$

We consider three possibilities for the weighting scheme in (3.29):

Discrete weighting in which the measure takes the form $v = \sum_{i=1}^N k_i \delta_{\alpha_i}$ for $N \geqslant 1$. This places positive mass $k_1, \ldots, k_N$ at the ordered values $\alpha_1 < \cdots < \alpha_N$. In this case $W_{v,t}$ in (3.29) becomes

$$W_{v,t} = \sum_{i=1}^N k_i I_{\{P_t > \alpha_i\}}\,. \tag{3.31}$$

In the case where $N = 1$ and $k_1 = 1$, we obtain $W_{v,t} = I_{\{P_t > \alpha\}}$, so that $(W_{v,t})$ is a series of iid Bernoulli$(1-\alpha)$ variables under the null hypothesis (3.30). In

the general case $(W_{v,t})$ is a series of iid ordered multinomial random variables taking the values $q_0 < q_1 < \cdots < q_N$ where $q_0 = 0$ and $q_k = \sum_{i=1}^{k} k_i$ for $k = 1, \ldots, N$. Under the null hypothesis (3.30) the distribution of $(W_{v,t})$ satisfies

$$\mathrm{P}\left(W_{v,t} = q_i\right) = \alpha_{i+1} - \alpha_i, \quad i \in \{0, 1, \ldots, N\}\,, \tag{3.32}$$

where $\alpha_0 = 0$ and $\alpha_{N+1} = 1$.

Continuous weighting in which the measure takes the form $\mathrm{d}v(u) = g(u)\,\mathrm{d}u$ on the interval $[\alpha_1, \alpha_2] \subset [0, 1]$, where the function g satisfies

Assumption 3.1. (i) $g(u) = 0$ for $u \notin [\alpha_1, \alpha_2]$, (ii) g is continuous and (iii) $g(u) > 0$ for $u \in (\alpha_1, \alpha_2)$.

For notation simplicity, we denote $W_{v,t}$ in (3.29) under the continuous weighting case by

$$W_{g,t} = \int_{\alpha_1}^{\alpha_2} g(u) I_{\{P_t > u\}} \mathrm{d}u\,. \tag{3.33}$$

Writing $G(\alpha) = \int_0^{\alpha} g(u)\mathrm{d}u$ for the integral of g we can derive the expression

$$W_{g,t} = I_{\{P_t > \alpha_1\}} G(\min(P_t, \alpha_2))\,. \tag{3.34}$$

When $g(u) = (\alpha_2 - \alpha_1)^{-1}$ is the uniform weighting function, we obtain

$$W_{g,t} = I_{\{P_t > \alpha_1\}} \frac{\min(P_t, \alpha_2) - \alpha_1}{\alpha_2 - \alpha_1}\,. \tag{3.35}$$

Combined discrete and continuous weighting. We can also consider a weighting scheme that is given by the sum of a discrete weighting and a continuous weighting scheme.

3.4.2 Some useful results for continuous weighting of realized p-value

In this section we derive some useful results for the continuous weighting scheme described by (3.33) where g satisfy Assumption 3.1 and G is its integral. An alternative expression to (3.34) is

$$W_{g,t} = G(\min(\alpha_2, \max(P_t, \alpha_1)))\,. \tag{3.36}$$

Suppose we denote the truncated realized PIT values by

$$P_t^* = \min(\alpha_2, \max(P_t, \alpha_1))\,. \tag{3.37}$$

(3.36) gives the useful insight that $W_{g,t}$ is a strictly increasing and continuous function of P_t^*.

The following results will be useful for computing moments and covariances of $W_{g,t}$ for different choices of weighting function g, which will be required for constructing tests later.

Proposition 3.2. Let

$$W_{g,t} = \int_{\alpha_1}^{\alpha_2} g(u) I_{\{P_t > u\}} \mathrm{d}u\,, \tag{3.38}$$

where g satisfies Assumption 3.1. We define $W_{g^*,t} = W_{g_i,t} W_{g_j,t}$, where g^*, g_i and g_j satisfy Assumption 3.1. We can calculate

$$g^*(u) = g_i(u)\, G_j(u) + g_j(u)\, G_i(u)\,. \tag{3.39}$$

Proof.

$$\begin{aligned}
W_{g_i,t} W_{g_j,t} &= \left(\int_{\alpha_1}^{\alpha_2} g_i(u) I_{\{P_t > u\}} \mathrm{d}u \right) \left(\int_{\alpha_1}^{\alpha_2} g_j(v) I_{\{P_t > v\}} \mathrm{d}v \right) \\
&= \int_{u=\alpha_1}^{\alpha_2} \int_{v=\alpha_1}^{\alpha_2} g_i(u) g_j(v) I_{\{P_t > u\}} I_{\{P_t > v\}} \mathrm{d}v \mathrm{d}u \\
&= \int_{u=\alpha_1}^{\alpha_2} \int_{v=\alpha_1}^{\alpha_2} g_i(u) g_j(v) I_{\{P_t > \max\{u,v\}\}} \mathrm{d}v \mathrm{d}u \\
&= \int_{u=\alpha_1}^{\alpha_2} \int_{v=\alpha_1}^{u} g_i(u) g_j(v) I_{\{P_t > u\}} \mathrm{d}v \mathrm{d}u + \int_{u=\alpha_1}^{\alpha_2} \int_{v=u}^{\alpha_2} g_i(u) g_j(v) I_{\{P_t > v\}} \mathrm{d}v \mathrm{d}u \\
&= \int_{u=\alpha_1}^{\alpha_2} g_i(u) \left(\int_{v=\alpha_1}^{u} g_j(v) \mathrm{d}v \right) I_{\{P_t > u\}} \mathrm{d}u + \int_{v=\alpha_1}^{\alpha_2} g_j(v) \left(\int_{u=\alpha_1}^{v} g_i(u) \mathrm{d}u \right) I_{\{P_t > v\}} \mathrm{d}v \\
&= \int_{u=\alpha_1}^{\alpha_2} g_i(u)\, G_j(u) \mathrm{d}u + \int_{v=\alpha_1}^{\alpha_2} g_j(v)\, G_i(v) \mathrm{d}v\,.
\end{aligned} \tag{3.40}$$

□

Proposition 3.3. Let

$$W_{g,t}^k = \int_{\alpha_1}^{\alpha_2} g_k^*(u) I_{\{P_t > u\}}\, \mathrm{d}u\,, \tag{3.41}$$

where g and g_k^* satisfy Assumption 3.1. For $k \geqslant 1$,

$$g_k^*(u) = k\, g(u)\, G(u)^{k-1}\,. \tag{3.42}$$

Proof. We will prove by induction. The result is true for $k = 2$ by Proposition 3.2. Assuming that the result is true for some k,

$$\begin{aligned} W_{g,t}^{k+1} &= \int_{\alpha_1}^{\alpha_2} g_{k+1}^*(u) I_{\{P_t > u\}} \, \mathrm{d}u \\ &= W_{g,t} \, W_{g,t}^k \, . \end{aligned} \tag{3.43}$$

Using the results in Proposition 3.2,

$$\begin{aligned} g_{k+1}^*(u) &= g(u) \, G_k^*(u) + g_k^*(u) \, G(u) \\ &= g(u) \, G(u)^k + k \, g(u) \, G(u)^k \\ &= (k+1) \, g(u) \, G(u)^k \, . \end{aligned} \tag{3.44}$$

□

To compute the p-centered moment $E((W_{g,t} - \mu_g)^p)$, where $\mu_g = \mathrm{E}\,(W_{g,t})$ calculated under H_0 in (3.30), one way would be use expand the terms and use the results in Proposition 3.3. A more convenient method would be to use the following lemma:

Lemma 3.4. For a set of weight functions $g_1, \ldots, g_k$, where g_i for $i = 1, \ldots, k$ satisfy Assumption 3.1, then using (3.36) we can calculate

$$\begin{aligned} \mathrm{E}\left(\prod_{i=1}^{k} (W_{g_i,t} - \mu_{g_i})^{p_i} \right) &= \alpha_1 \prod_{i=1}^{k} (-\mu_{g_i})^{p_i} + (1 - \alpha_2) \prod_{i=1}^{k} (G_i(\alpha_2) - \mu_{g_i})^{p_i} + \\ &\quad \int_{\alpha_1}^{\alpha_2} \prod_{i=1}^{k} (G_i(u) - \mu_{g_i})^{p_i} \mathrm{d}u \, , \end{aligned} \tag{3.45}$$

where $\mu_{g_i} = \mathrm{E}\,(W_{g_i,t})$ calculated under H_0 in (3.30).

3.4.3 Spectral test

We denote by μ_v and σ_v^2 the mean and variance of $W_{v,t}$ calculated under H_0 in (3.30), where $W_{v,t}$ is defined in (3.29). We can construct a test based on CLT, where the Z-test statistic is given by

$$Z_v = \sqrt{\frac{n}{\sigma_v^2}} \, (\hat{\mu}_v - \mu_v) \, , \tag{3.46}$$

where $\hat{\mu}_v = \frac{1}{n} \sum_{t=1}^{n} W_{v,t}$. This form of test has been proposed by Costanzino & Curran (2015), which we will refer to as the spectral test.

Discrete weighting. In the case where $W_{v,t}$ is defined by (3.31), under the null hypothesis (3.30), we can calculate

$$\mu_v = \sum_{i=1}^{N} k_i(1-\alpha_j)\,, \tag{3.47}$$

$$\mu_{v,2} = \sum_{i=1}^{N}\sum_{j=1}^{N} k_i k_j(1-\max(\alpha_i,\alpha_j))\,, \tag{3.48}$$

and the variance is given by $\sigma_v^2 = \mu_{v,2} - \mu_v^2$.

Continuous weighting. In the case where $W_{v,t}$ is defined by (3.33), under the null hypothesis (3.30), we can calculate

$$\mu_v = \int_{\alpha_1}^{\alpha_2} g(u)(1-u)\,\mathrm{d}u \tag{3.49}$$

$$\mu_{v,2} = \int_{\alpha_1}^{\alpha_2} 2\,g(u)\,G(u)(1-u)\,\mathrm{d}u\,, \tag{3.50}$$

where the second equation is obtain using the results in Proposition 3.2, and the variance is given by $\sigma_v^2 = \mu_{v,2} - \mu_v^2$.

For the simulation studies in the sequel, we will focus on the continuous weighting case, as the results for the discrete weighting case is similar in terms of size and power performance. We consider the following choice of weight function for the spectral test:

Uniform weight. $g(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}}$ as considered in Costanzino & Curran (2015) and Costanzino & Curran (2016) with $G(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}} u$ and

$$\mu_v = \left[-\frac{1}{2}(1-u)^2\right]_{u=\alpha_1}^{\alpha_2}, \tag{3.51}$$

$$\mu_{v,2} = \left[(1+\alpha_1)u^2 - 2\alpha_1 u - \frac{2}{3}u^3\right]_{u=\alpha_1}^{\alpha_2}. \tag{3.52}$$

Linear weight. $g(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}}(u-\alpha_1)$ with $G(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}}\frac{1}{2}(u-\alpha_1)^2$ and

$$\mu_v = \left[\frac{1}{2}u^2(1+\alpha_1) - \alpha_1 u - \frac{1}{3}u^3\right]_{u=\alpha_1}^{\alpha_2}, \tag{3.53}$$

$$\mu_{v,2} = \left[\frac{1}{4}(1-\alpha_1)(u-\alpha_1)^4 - \frac{1}{5}(u-\alpha_1)^5\right]_{u=\alpha_1}^{\alpha_2}. \tag{3.54}$$

Exponential weight. $g(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}} e^{k(u-\alpha_1)}$ with $G(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}} \frac{1}{k} e^{k(u-\alpha_1)}$ and

$$\mu_v = \left[\frac{1}{k} \left(e^{k(u-\alpha_1)} - 1 \right) \right]_{u=\alpha_1}^{\alpha_2}, \tag{3.55}$$

$$\mu_{v,2} = \left[\frac{-e^{k(u-\alpha_1)}}{k^3} \left(e^{k(u-\alpha_1)} \left(k(u-1) - \frac{1}{2} \right) + 2 + k(2-2u) \right) \right]_{u=\alpha_1}^{\alpha_2}. \tag{3.56}$$

3.4.4 Bispectral test

In this section we propose a new test that extends the idea of the spectral test. Consider $\boldsymbol{W}_{\boldsymbol{v},t} = (W_{v_1,t}, W_{v_2,t})^T$ where the $W_{v_i,t}$ is defined in (3.29). The null hypothesis (3.30) needs to be generalized to

$$H_0 : \quad (\boldsymbol{W}_{\boldsymbol{v},t}) \text{ is an iid series of vectors with df } F_{\boldsymbol{W}}, \tag{3.57}$$

where $F_{\boldsymbol{W}}$ denotes the distribution function of $(\boldsymbol{W}_{\boldsymbol{v},t})$ when P_t is uniform. Under the null hypothesis (3.57), we can apply the multivariate version of the CLT to construct the Z-test statistic

$$\boldsymbol{Z}_{\boldsymbol{v}} = \sqrt{n}\, \Sigma_{\boldsymbol{v}}^{-1/2} \left(\overline{\boldsymbol{W}} - \boldsymbol{\mu}_{\boldsymbol{v}} \right), \tag{3.58}$$

where $\overline{\boldsymbol{W}} = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{W}_{\boldsymbol{v},t}$, and $\boldsymbol{\mu}_{\boldsymbol{v}} = (\mu_{v_1}, \mu_{v_2})^T$ is the mean vector and

$$\Sigma_{\boldsymbol{v}} = \begin{pmatrix} \sigma_{v_1}^2 & \sigma_{v_1,v_2} \\ \sigma_{v_1,v_2} & \sigma_{v_2}^2 \end{pmatrix}, \tag{3.59}$$

is the covariance matrix. We have already discussed the calculation of the mean and variance of $W_{v_i,t}$ in Section 3.4.3.

Discrete weighting. In the case where $W_{v_i,t}$ is defined by (3.31), with $\nu_1 = \sum_{i=1}^{N} a_i \delta_{\alpha_i}$ and $\nu_2 = \sum_{i=1}^{N} b_i \delta_{\alpha_i}$, with corresponding $W_{v_1,t} = \sum_{i=1}^{N} a_i I_{\{P_t > \alpha_i\}}$ and $W_{v_2,t} = \sum_{i=1}^{N} b_i I_{\{P_t > \alpha_i\}}$, under the null hypothesis (3.57), we can calculate the off-diagonal element of the matrix $\Sigma_{\boldsymbol{v}}$ with

$$\sigma_{v_1,v_2} = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i b_j \big(1 - \max(\alpha_i, \alpha_j) - (1-\alpha_i)(1-\alpha_j) \big). \tag{3.60}$$

Continuous weighting. In the case where $W_{v_i,t}$ is defined by (3.33), with $\mathrm{d}v_i(u) = g_i(u)\,\mathrm{d}u$, where the weight functions g_1 and g_2 satisfy Assumption 3.1, under the null hypothesis (3.57), the off-diagonal element of the matrix $\Sigma_{\boldsymbol{v}}$ is $\sigma_{v_1,v_2} = \mathrm{E}\left(W_{v_1,t}\, W_{v_2,t}\right) - \mu_{v_1}\mu_{v_2}$ can be calculated using Proposition 3.2.

Under the null hypothesis (3.57), we assume that $\boldsymbol{Z_v} \sim N_2(\mathbf{0}, I_2)$, where I_2 denotes the 2×2 identity matrix. For a two sided test, we assume that the test statistic

$$S_{\boldsymbol{v}} = n\left(\overline{\boldsymbol{W}} - \boldsymbol{\mu_v}\right)^T \Sigma_{\boldsymbol{v}}^{-1}\left(\overline{\boldsymbol{W}} - \boldsymbol{\mu_v}\right) \sim \chi_2^2 . \tag{3.61}$$

The bispectral test can be extended naturally to a k-spectral test, in which $(\boldsymbol{W}_{\boldsymbol{v},t})$ is a series of vectors with dimension k.

For the simulation studies in the sequel, we will focus on the continuous weighting case with $\mathrm{d}v_i(u) = g_i(u)\,\mathrm{d}u$, where we consider the following combinations of weight functions:

Uniform-Linear weight. $g_1(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}}$, $g_2(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}}(u - \alpha_1)$, and

$$\mathrm{E}\left(W_{v_1,t} W_{v_2,t}\right) = \left[\frac{1}{2}(1-\alpha_1)(u-\alpha_1)^3 - \frac{3}{8}(u-\alpha_1)^4\right]_{u=\alpha_1}^{\alpha_2} . \tag{3.62}$$

Uniform-Exponential weight. $g_1(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}}$, $g_2(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}} k e^{ku}$, and

$$\mathrm{E}\left(W_{v_1,t} W_{v_2,t}\right) = \left[\frac{e^{k(u-\alpha_1)}}{k^3}\left((u-1)(\alpha_1-u)k^2 + (u-\alpha_1)k - 1\right) + \frac{u(u-2)}{2k}\right]_{u=\alpha_1}^{\alpha_2} . \tag{3.63}$$

Linear-Exponential weight. $g_1(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}}(u - \alpha_1)$, $g_2(u) = I_{\{\alpha_1 \leqslant u \leqslant \alpha_2\}} k e^{ku}$, and

$$\mathrm{E}\left(W_{v_1,t} W_{v_2,t}\right) = \left[\frac{e^{k(u-\alpha_1)}}{k^4}\left(\frac{1}{2}(1-u)(\alpha_1-u)^2k^3 + \frac{1}{2}(\alpha_1-u)^2k^2 + (\alpha_1-u)k + 1\right) - \frac{u}{2k}\left((\alpha_1+1)u - \frac{2}{3}u^2 - 2\alpha_1\right)\right]_{u=\alpha_1}^{\alpha_2} . \tag{3.64}$$

3.4.5 One-sided spectral and bispectral test

For regulatory purposes, it may be important to construct a one-sided test, since the regulators are more concerned if financial institutions do not set aside enough capital.

We first focus on the one-sided spectral test. Assuming that the weight measure v in (3.29) is always positive, when the VaR exception rate at the tail is larger than expected, which occurs when the forecast distribution underestimates the tail of the underlying distribution of losses, $W_{v,t}$ will be large as well. Hence, it would be sensible to test the hypothesis

$$H_0 : \mathrm{E}\,(W_{v,t}) \leqslant \mu_v \quad \text{vs.} \quad H_1 : \mathrm{E}\,(W_{v,t}) > \mu_v\,. \tag{3.65}$$

To obtain a one-sided Z-test with approximate size κ, we reject H_0 in (3.65) when $Z_v > \Phi^{-1}\,(1-\kappa)$, where Z_v is the spectral test statistic in (3.46).

Similar arguments hold for the bispectral case, where we propose to use a sequential rejection approach. Using Cholesky decomposition we can construct the vector

$$\begin{pmatrix} \tilde{Z}_{v_1} \\ \tilde{Z}_{v_2} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{n}(\overline{W}_{v_1}-\mu_{v_1})}{\sigma_{v_1}} \\ \frac{\sqrt{n}(\overline{W}_{v_2}-\mu_{v_2})}{\sigma_{v_2}\sqrt{1-\rho^2}} - \frac{\rho\,\tilde{Z}_{v_1}}{\sqrt{1-\rho^2}} \end{pmatrix}, \tag{3.66}$$

which is asymptotically $N_2(\mathbf{0}, I_2)$ distributed, where $\overline{W}_{v_i} = \frac{1}{n}\sum_{t=1}^{n} W_{v_i,t}$, and $\rho = \frac{\sigma_{v_1,v_2}}{\sqrt{\sigma_{v_1}\sigma_{v_2}}}$ is the correlation between $W_{v_1,t}$ and $W_{v_2,t}$. To construct a test of approximately size κ, we first test the hypothesis

$$H_0 : \mathrm{E}\,(W_{v_1,t}) \leqslant \mu_{v_1} \quad \text{vs.} \quad H_1 : \mathrm{E}\,(W_{v_1,t}) > \mu_{v_1}\,. \tag{3.67}$$

We accept H_0 in (3.67) if $\tilde{Z}_{v_1} \leqslant \Phi^{-1}\left(1-\frac{\kappa}{2}\right)$. Given that we accept H_0, for a given $\tilde{Z}_{v_1}$, we proceed to test the hypothesis

$$H_0 : \mathrm{E}\,(W_{v_2,t}) \leqslant \mu_{v_2} \quad \text{vs.} \quad H_1 : \mathrm{E}\,(W_{v_2,t}) > \mu_{v_2}\,, \tag{3.68}$$

where we accept H_0 in (3.68) if $\tilde{Z}_{v_2} \leqslant \Phi^{-1}\left(1-\frac{\kappa}{2}\right)$. This is the same as applying the Bonferroni (1936) multiple test procedure to test the global hypothesis

$$\begin{aligned} H_0 :&\quad \mathrm{E}\,(W_{v_1,t}) \leqslant \mu_{v_1} \quad \text{and} \quad \mathrm{E}\,(W_{v_2,t}) \leqslant \mu_{v_2}\,, \\ H_1 :&\quad \mathrm{E}\,(W_{v_1,t}) > \mu_{v_1} \quad \text{or} \quad \mathrm{E}\,(W_{v_2,t}) > \mu_{v_2}\,, \end{aligned} \tag{3.69}$$

and to obtain a test of approximately size κ, we reject H_0 in (3.69) when

$$\min\left(1-\Phi(\tilde{Z}_{v_1}), 1-\Phi(\tilde{Z}_{v_2})\right) < \kappa/2\,. \tag{3.70}$$

See Hommel et al. (2011) for a review of several other multiple test procedures. We have used the Cholesky decomposition to construct the vector in (3.66) to improve the size performance of the Bonferroni test.

3.5 Truncated probitnormal score test

In this section we introduce the truncated probitnormal score test, which is a special case of the bispectral test, where the weight measure v_1 and v_2 is given by the sum of a discrete weighting with point mass at α_1 and α_2, and continuous weight functions g_1 and g_2 which satisfy Assumption 3.1.

In the probitnormal score test we assume that the underlying distribution of P_t to be probitnormal, so that $\Phi^{-1}(P_t) \sim N(\mu, \sigma^2)$. We denote the parameter vector as $\boldsymbol{\theta} = (\mu, \sigma)^T$, and the distribution function and density of P_t are given by

$$F_P(p \mid \boldsymbol{\theta}) = \Phi\left(\frac{\Phi^{-1}(p)-\mu}{\sigma}\right), \quad f_P(p \mid \boldsymbol{\theta}) = \frac{\phi\left(\frac{\Phi^{-1}(p)-\mu}{\sigma}\right)}{\phi(\Phi^{-1}(p))\sigma}\,, \quad p \in [0,1]\,. \tag{3.71}$$

The above construction provides a flexible family in which we can test for a uniform distribution corresponding to $\boldsymbol{\theta} = \boldsymbol{\theta}_0 = (0,1)^T$.

We may be interested in testing the tail of the distribution only. We denote the truncated realized PIT values by $P_t^* = \min(\alpha_2, \max(P_t, \alpha_1))$, with corresponding likelihood function

$$L(\boldsymbol{\theta} \mid P_t^*) = \begin{cases} F_P(\alpha_1 \mid \boldsymbol{\theta}) & P_t^* = \alpha_1\,, \\ f_P(P_t^* \mid \boldsymbol{\theta}) & \alpha_1 < P_t^* < \alpha_2\,, \\ 1 - F_P(\alpha_2 \mid \boldsymbol{\theta}) & P_t^* = \alpha_2\,. \end{cases} \tag{3.72}$$

We denote the observed score vector for P_t^* by

$$\boldsymbol{S}_t(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial\mu}\log L(\boldsymbol{\theta} \mid P_t^*), \frac{\partial}{\partial\sigma}\log L(\boldsymbol{\theta} \mid P_t^*)\right)^T, \tag{3.73}$$

and the expected Fisher information matrix by $I(\boldsymbol{\theta})$, which is given by

$$I(\boldsymbol{\theta})_{ij} = -\mathrm{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln L(\boldsymbol{\theta} \mid P_t)\right). \tag{3.74}$$

For the truncated probitnormal distribution the matrix of second derivatives is given in Appendix A (equations (A.3), (A.4) and (A.5)).

Under the null hypothesis that (P_t) is a series of iid standard uniform variables, We can construct the score test statistic

$$\boldsymbol{Z_v} = \sqrt{n} I(\boldsymbol{\theta}_0)^{-1/2} \overline{\boldsymbol{S}}(\boldsymbol{\theta}_0)\,, \tag{3.75}$$

which is asymptotically $N_2\big(\mathbf{0}, I_2\big)$ distributed, where $\overline{\boldsymbol{S}}(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{S}_t(\boldsymbol{\theta}_0)$ is the mean of the observed score vectors evaluated under the null, and $I(\boldsymbol{\theta}_0)^{-1/2}$ is the inverse of the cholesky decomposition of $I(\boldsymbol{\theta}_0)$, and I_2 is the 2×2 identity matrix. For a two-sided test, we assume that

$$n\overline{\boldsymbol{S}}(\boldsymbol{\theta}_0)^T I(\boldsymbol{\theta}_0)^{-1} \overline{\boldsymbol{S}}(\boldsymbol{\theta}_0) \sim \chi^2_2\,. \tag{3.76}$$

In the following result we show that this is a bispectral test with the structure in (3.61).

Proposition 3.5. $\boldsymbol{S}_t(\boldsymbol{\theta}_0) = \boldsymbol{W}_{\boldsymbol{v},t} - \boldsymbol{\mu}_{\boldsymbol{v}}$, almost surely, where

$$W_{v_i,t} = k_{i,1} I_{\{P_t > \alpha_1\}} + k_{i,2} I_{\{P_t > \alpha_2\}} + \int_{\alpha_1}^{\alpha_2} g_i(u) I_{\{P_t > u\}} \mathrm{d}u\,, \tag{3.77}$$

and the weight function g_i satisfies Assumption 3.1.

Proof. Computing the score statistic and evaluating it at $\boldsymbol{\theta}_0 = (0,1)^T$ yields

$$\boldsymbol{S}_t(\boldsymbol{\theta}_0) = \begin{cases} \begin{pmatrix} -\phi(\Phi^{-1}(\alpha_1))/\alpha_1 \\ -\phi(\Phi^{-1}(\alpha_1))\Phi^{-1}(\alpha_1)/\alpha_1 \end{pmatrix} & P_t^* = \alpha_1\,, \\ \begin{pmatrix} \Phi^{-1}(P_t^*) \\ \Phi^{-1}(P_t^*)^2 - 1 \end{pmatrix} & \alpha_1 < P_t^* < \alpha_2\,, \\ \begin{pmatrix} \phi(\Phi^{-1}(\alpha_2))/(1-\alpha_2) \\ \phi(\Phi^{-1}(\alpha_2))\Phi^{-1}(\alpha_2)/(1-\alpha_2) \end{pmatrix} & P_t^* = \alpha_2\,. \end{cases} \tag{3.78}$$

The jumps at α_1 and α_2 are given by

$$\begin{aligned} k_{1,1} &= \Phi^{-1}(\alpha_1) + \tfrac{\phi(\Phi^{-1}(\alpha_1))}{\alpha_1}\,, & k_{1,2} &= \tfrac{\phi(\Phi^{-1}(\alpha_2))}{1-\alpha_2} - \Phi^{-1}(\alpha_2)\,, \\ k_{2,1} &= \Phi^{-1}(\alpha_1)^2 - 1 + \tfrac{\phi(\Phi^{-1}(\alpha_1))\Phi^{-1}(\alpha_1)}{\alpha_1}\,, & k_{2,2} &= \tfrac{\phi(\Phi^{-1}(\alpha_2))\Phi^{-1}(\alpha_2)}{1-\alpha_2} - \Phi^{-1}(\alpha_2)^2 + 1\,. \end{aligned} \tag{3.79}$$

The weight measures can be obtained by differentiating $\boldsymbol{S}_t(\boldsymbol{\theta}_0)$ with respect to P_t on $[\alpha_1, \alpha_2]$, and are given by

$$v_1(u) = I_{\{\alpha_1<P_t<\alpha_2\}} g_1(u) + \delta_{\alpha_1} k_{1,1} + \delta_{\alpha_2} k_{1,2}\,, \tag{3.80}$$

$$v_2(u) = I_{\{\alpha_1<P_t<\alpha_2\}} g_2(u) + \delta_{\alpha_1} k_{2,1} + \delta_{\alpha_2} k_{2,2}\,, \tag{3.81}$$

where $g_1(u) = \frac{1}{\phi(\Phi^{-1}(u))}$ and $g_2(u) = \frac{2\Phi^{-1}(u)}{\phi(\Phi^{-1}(u))}$. The mean of $\boldsymbol{W}_{\boldsymbol{v},t}$ is given by

$$\mu_{v_1} = \int_{\alpha_1}^{\alpha_2} g_1(u)(1-u)\,\mathrm{d}u + (1-\alpha_1)k_{1,1} + (1-\alpha_2)k_{1,2}\,, \tag{3.82}$$

$$\mu_{v_2} = \int_{\alpha_1}^{\alpha_2} g_2(u)(1-u)\,\mathrm{d}u + (1-\alpha_1)k_{2,1} + (1-\alpha_2)k_{2,2}\,. \tag{3.83}$$

Note that since $\boldsymbol{W}_{\boldsymbol{v},t} = \boldsymbol{S}_t(\boldsymbol{\theta}_0) + \boldsymbol{\mu}_{\boldsymbol{v}}$, and $\boldsymbol{W}_{\boldsymbol{v},t} = 0$ when $P_t^* = \alpha_1$, we must have

$$\mu_{v_1} = \frac{\phi(\Phi^{-1}(\alpha_1))}{\alpha_1}\,, \tag{3.84}$$

$$\mu_{v_2} = \frac{\phi(\Phi^{-1}(\alpha_1))\Phi^{-1}(\alpha_1)}{\alpha_1}\,. \tag{3.85}$$

□

Note that the weighting functions $g_i(u)$ approach infinity as u approaches one, so a very large weight is applied to the largest realized PIT values. The covariance matrix $\Sigma_{\boldsymbol{v}} = \mathrm{cov}(\boldsymbol{W}_{\boldsymbol{v},t}) = I(\boldsymbol{\theta}_0)$ is given in Appendix A (equations (A.6), (A.7) and (A.8)).

The following lemma will be useful for performing size correction on the martingale difference test for testing serial independence later.

Lemma 3.6. We denote by $\boldsymbol{S}_t(\boldsymbol{\theta}_0)_u = (S_{1,t}(\boldsymbol{\theta}_0)_u, S_{2,t}(\boldsymbol{\theta}_0)_u)^T$ the score vector in (3.78) evaluated at $P_t^* = u$. For a set of weight measures $v_1, \ldots, v_k$, where v_1 and v_2 are the probitnormal weight measures given by (3.80) and (3.81), and $\mathrm{d}v_i(u) = g_i(u)\,\mathrm{d}u$ with g_i satisfying Assumption 3.1 for $i = 3, \ldots, k$, we can calculate

$$\mathrm{E}\left(\prod_{i=1}^{k}(W_{v_i,t} - \mu_{v_i})^{p_i}\right) = \alpha_1 \prod_{i=1}^{k}(-\mu_{v_i})^{p_i} + (1-\alpha_2)\prod_{i=1}^{k} f_i(\alpha_2)^{p_i} + \int_{\alpha_1}^{\alpha_2}\prod_{i=1}^{k} f_i(u)^{p_i}\mathrm{d}u\,, \tag{3.86}$$

where $f_i(u) = S_{i,t}(\boldsymbol{\theta}_0)_u$ for $i = 1, 2$ and $f_i(u) = G_i(u) - \mu_{v_i}$ for $i = 3, \ldots, k$, and $\mu_{v_i} = \mathrm{E}\,(W_{v_i,t})$ is calculated under H_0 in (3.30).

3.5.1 Truncated Probitnormal LRT

An alternative to the probitnormal score test is the truncated probitnormal LRT, where we want to test the hypothesis that $\mu = 0$ and $\sigma = 1$ in the context of the truncated probitnormal model described by the likelihood in (3.72). We test the null hypothesis that $(W_{v,t})$ have the distribution implied by the independence and uniformity of (P_t) against the alternative that they have the distribution implied by (P_t) having a probitnormal distribution.

Suppose we consider the continuous weighting case where $W_{v,t}$ is defined by (3.33). In the one-sided case where $\alpha_1 = \alpha$ and $\alpha_2 = 1$, and $g(u) = (1-\alpha)^{-1}$, this test is identical to the test proposed by Berkowitz (2001). In the paper $Z_t = \Phi^{-1}(P_t^*)$ is modeled by a normal distribution truncated to $[\Phi^{-1}(\alpha_1), \infty)$, where $P_t^* = \max(P_t, \alpha_1)$ and

$$\mathrm{P}\left(Z_t \leqslant z\right) = \Phi\left(\frac{z-\mu}{\sigma}\right), \quad z \geqslant \Phi^{-1}(\alpha)\,. \tag{3.87}$$

Recall from (3.35) that

$$\begin{aligned} W_{g,t} &= \frac{P_t^* - \alpha}{1-\alpha} \\ &= \frac{\Phi(Z_t) - \alpha}{1-\alpha} \end{aligned} \tag{3.88}$$

has uniform density $(1-\alpha)^{-1}$ on (0,1). The Berkowitz model (3.87) is equivalent to a model where

$$\mathrm{P}\left(W_{g,t} \leqslant w\right) = \Phi\left(\frac{\Phi^{-1}(\alpha + w(1-\alpha)) - \mu}{\sigma}\right), \quad w \in [0,1)\,, \tag{3.89}$$

and we test for $\mu = 0$ and $\sigma = 1$.

Note that the multinomial LRT model (3.22) that we consider, with $\alpha_i = \alpha + \frac{i-1}{N}(1-\alpha)$, for $i = 1, \ldots, N$, coincides with the Berkowitz test in the limit as the number of levels N goes to infinity. To see this, recall that

$$\mathrm{P}\left(C_t \leqslant i\right) = \Phi\left(\frac{\Phi^{-1}(\alpha + \frac{i}{N}(1-\alpha)) - \mu}{\sigma}\right), \quad i = 0, \ldots, N\,, \tag{3.90}$$

which is the natural discrete counterpart of the continuous model in (3.89), where C_t is defined in (3.8), and we test for $\mu = 0$ and $\sigma = 1$.

We make the following notes regarding the role of the weight measure v in the construction of a likelihood ratio test. The weight measure v will determine which

distribution should be used to construct a LRT for $W_{v,t}$. For example, when the measure v is standard uniform, F_W in (3.30) is standard uniform, so that the appropriate distributions to be used in the contruction of LRT are distributions which nest the standard uniform distribution, such as the probitnormal distribution and beta distribution. Similarly, in the continuous case where the weight measure is g_1 in Proposition 3.5, F_W is standard normal, and hence the appropriate distribution to be used is the normal distribution. This corresponds to the standard Berkowitz test.

For a given distribution, the weight measure v serves as the mapping function, and plays no role in the rejection rate of the LRT. The easiest way to understand this is to consider the discrete weight measure $v_i = k_i\delta_\alpha$, with $k_i > 0$, so that when P_t is standard uniform, $W_{v_i,t}$ takes the value zero with probability α and k_i with probability $1-\alpha$. Suppose we construct a LRT for $W_{v_i,t}$, where we model $W_{v_i,t}$ to take the value zero with probability θ and k_i with probability $1-\theta$. The likelihood of $W_{v_i,t}$ is then given by $I_{\{W_{v_i,t}=0\}}\,\theta + I_{\{W_{v_i,t}=k_i\}}\,(1-\theta)$, which is the same for all i, and does not depend on the choice of k_i.

For a chosen weight measure v, the rejection rate of the constructed LRT depends on the distribution used to model $W_{v,t}$. We know from the results in Chapter 2 that the LRT is derived from the two-sided score test with some additional approximations. Hence, we would expected the truncated probitnormal LRT to give similar results to the two-sided truncated probitnormal score test. More generally, suppose we construct the LRT for $W_{v,t}$ based on some distribution with a k-dimensional parameter vector. The results for the LRT will be similar to those of a k-spectral test, where the weight measures can be obtained by differentiating the score vector of the distribution with respect to P_t.

3.5.2 One-sided probitnormal test

From Proposition 3.5. we know that the probitnormal score test is a special case of the bispectral test, where the weight measures v_1 and v_2 is positive in the range $[\alpha_1, \alpha_2]$. We denote by $\boldsymbol{S}_t(\boldsymbol{\theta}_0) = (S_{1,t}(\boldsymbol{\theta}_0), S_{2,t}(\boldsymbol{\theta}_0))^T$ the score vector for the truncated probitnormal distribution (3.73). For the probitnormal case, the vector (3.66) can be written as

$$\begin{pmatrix} \tilde{Z}_{v_1} \\ \tilde{Z}_{v_2} \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{n}\,\overline{S}_1(\boldsymbol{\theta}_0)}{\sigma_{v_1}} \\ \frac{\sqrt{n}\,\overline{S}_2(\boldsymbol{\theta}_0)}{\sigma_{v_2}\sqrt{1-\rho^2}} - \frac{\rho\,\tilde{Z}_{v_1}}{\sqrt{1-\rho^2}} \end{pmatrix}, \tag{3.91}$$

where we denoted $\overline{S}_i(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{t=1}^{n} S_{i,t}(\boldsymbol{\theta}_0)$, and the variances are given by $\sigma_{v_1}^2 = I(\boldsymbol{\theta}_0)_{1,1}$ and $\sigma_{v_2}^2 = I(\boldsymbol{\theta}_0)_{2,2}$, and the correlation by $\rho = \frac{I(\boldsymbol{\theta}_0)_{1,2}}{\sqrt{I(\boldsymbol{\theta}_0)_{1,1}\, I(\boldsymbol{\theta}_0)_{2,2}}}$. We can then construct a one-sided test based on Section 3.4.5, which we will refer to as the one-sided probitnormal score test.

In this section, we propose an alternative one-sided test based on the Wald test statistic, which we will refer to as the one-sided probitnormal Wald test. Table 1 shows the exception probability at the 99% level $\mathrm{E}\left(I_{\{P_t>0.99\}} \mid \mu, \sigma\right)$ for different parameter values when we assume the underlying distribution of P_t to be probitnormal, so that $\Phi^{-1}(P_t) \sim N(\mu, \sigma^2)$. We have coloured the cases where $\mathrm{E}\left(I_{\{P_t>0.99\}} \mid \mu, \sigma\right) \leqslant 0.01$ as green and the cases where $\mathrm{E}\left(I_{\{P_t>0.99\}} \mid \mu, \sigma\right) > 0.01$ as red.

$\sigma \mid \mu$	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5
0.75	0.01	0.01	0.02	0.04	0.06	0.10	0.15	0.23	0.34	0.51	0.74
0.80	0.02	0.03	0.05	0.08	0.12	0.18	0.27	0.39	0.57	0.80	1.12
0.85	0.04	0.07	0.10	0.15	0.22	0.31	0.44	0.62	0.86	1.17	1.58
0.90	0.08	0.12	0.18	0.25	0.35	0.49	0.67	0.91	1.22	1.62	2.12
0.95	0.15	0.20	0.28	0.39	0.53	0.72	0.96	1.26	1.65	2.13	2.73
1.00	0.24	0.32	0.43	0.58	0.76	1.00	1.30	1.67	2.14	2.70	3.39
1.05	0.36	0.47	0.62	0.81	1.04	1.34	1.70	2.14	2.68	3.33	4.10
1.10	0.51	0.66	0.85	1.08	1.37	1.72	2.15	2.66	3.27	4.00	4.84
1.15	0.70	0.89	1.12	1.40	1.74	2.15	2.64	3.22	3.90	4.70	5.61
1.20	0.92	1.15	1.43	1.76	2.16	2.63	3.18	3.82	4.57	5.42	6.40
1.25	1.19	1.46	1.78	2.16	2.61	3.14	3.75	4.45	5.25	6.17	7.20

Table 1: The exception probability at the 99% level $\mathrm{E}\left(I_{\{P_t>0.99\}} \mid \mu, \sigma\right)$ *for different parameter values, where we assumed that* $\Phi^{-1}(P_t) \sim N(\mu, \sigma^2)$. *Results are in percentage.*

The one-sided test for the VaR unconditional coverage hypothesis is

$$H_0 : \ \mathrm{E}\left(I_{\{P_t>0.99\}}\right) \leqslant 0.01 \ \ \text{vs.} \ \ H_1 : \mathrm{E}\left(I_{\{P_t>0.99\}}\right) > 0.01\,. \tag{3.92}$$

Assuming that P_t is probitnormal distributed, the exceedence rate is given by

$$\mathrm{E}\left(I_{\{P_t>0.99\}} \mid \mu, \sigma\right) = 1 - \Phi\left(\frac{\Phi^{-1}(0.99) - \mu}{\sigma}\right). \tag{3.93}$$

We define the function $\sigma_0(\mu)$ which solve the equation $\mathrm{E}\left(I_{\{P_t>0.99\}} \mid \mu, \sigma_0(\mu)\right) = 0.01$, i.e.

$$\sigma_0(\mu) = \frac{\Phi^{-1}(0.99) - \mu}{\Phi^{-1}(0.99)}\,, \tag{3.94}$$

then the hypothesis (3.92) can instead be written as

$$H_0 : \sigma \leqslant \sigma_0(\mu) \mid \mu \ \ \text{vs.} \ \ H_1 : \sigma > \sigma_0(\mu) \mid \mu\,, \tag{3.95}$$

where $\sigma_0(\mu)$ is defined in (3.94). Using the results in Section 2.3.1, the one-sided Wald test statistic is

$$Z_{\text{Wald}} \mid \hat{\mu} = \sqrt{n\, I(\boldsymbol{\theta}_0)_{2,2}}\big(\hat{\sigma} - \sigma_0(\hat{\mu})\big)\,, \tag{3.96}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the maximum likelihood estimators based on the truncated probitnormal likelihood (3.72). To obtain a test of approximately size κ, for a given $\hat{\mu}$ we reject H_0 in (3.95) when $Z_{\text{Wald}} > \Phi^{-1}(1-\kappa)$.

We now explore the size and power of the one-sided probitnormal tests. We simulate $(P_1, \ldots, P_n)$ using the equation $P_t = \Phi(\mu + \sigma Z_t)$, for varying μ and σ, where (Z_t) are iid standard normal variables. We then apply the one-sided probitnormal score test and the one-sided probitnormal Wald test with truncation levels $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$ to the simulated sequence (P_t). The resulting rejection rate based on 1000 simulations are shown in Table 2, where we set the size parameter $\kappa = 0.05$. We have applied the following colour coding: green indicates good results ($\leqslant 6\%$ for the size; $\geqslant 70\%$ for the power); red indicates poor results ($\geqslant 9\%$ for the size; $\leqslant 30\%$ for the power); dark red indicates very poor results ($\geqslant 12\%$ for the size; $\leqslant 10\%$ for the power). For the cases when H_0 in (3.95) is satisfied (the top left proportion of the black border), we apply the size colour coding, and for the cases when H_1 in (3.95) is satisfied (the bottom right proportion of the black border) the power colour coding.

We observe that the rejection rate pattern is roughly similar for both test. The size performance of the tests are poorest when μ is very negative, where the Wald test outperforms the score test, since it avoids the use of Bonferroni multiple test procedure. In contrast, the score test outperforms the Wald test in terms of power performance. Both the size and power improves with increasing sample size n. In conclusion, if the sample size n is relatively small, we should use the one-sided Wald test to ensure the size is reliable, otherwise we should use the one-sided score test.

One-sided probitnormal score test

n	250											500										
$\sigma \mid \mu$	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5
0.75	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.3	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	2.5
0.80	0.0	0.0	0.0	0.1	0.1	0.2	0.1	0.4	0.8	4.5	15.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	2.6	22.7
0.85	0.1	0.1	0.1	0.1	0.3	0.4	0.8	1.5	4.8	15.4	41.5	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.4	3.3	22.5	65.3
0.90	0.1	0.1	0.2	0.7	0.9	1.1	2.3	5.4	15.3	39.2	72.0	0.0	0.0	0.1	0.1	0.4	0.2	1.3	4.1	21.2	59.4	92.7
0.95	0.2	0.7	1.2	1.5	1.8	3.3	6.9	15.9	37.5	68.7	89.3	0.1	0.2	0.3	0.4	1.2	1.6	5.1	20.7	56.1	90.4	99.1
1.00	1.3	1.7	2.2	2.9	3.9	8.5	17.7	36.7	64.8	86.4	96.8	0.5	0.8	1.3	1.7	2.3	6.0	21.1	54.0	87.4	98.6	99.9
1.05	2.0	2.7	3.5	5.3	9.7	19.5	36.8	62.1	83.4	95.1	99.3	1.2	2.0	2.8	3.6	8.5	22.1	52.2	85.2	97.9	99.9	100.0
1.10	3.3	4.1	6.6	12.1	20.8	36.9	59.5	81.1	93.9	98.9	99.7	2.8	3.6	5.7	11.3	24.3	51.9	83.0	96.9	99.8	99.9	100.0
1.15	5.6	8.0	13.9	23.3	37.9	57.5	79.6	92.7	98.0	99.5	99.9	5.9	8.2	13.8	26.6	52.7	80.7	96.0	99.3	99.9	100.0	100.0
1.20	10.6	15.9	24.9	38.6	57.0	78.2	91.9	96.9	99.5	99.9	100.0	11.3	17.6	29.2	53.2	80.1	94.8	99.2	99.9	100.0	100.0	100.0
1.25	18.4	26.7	39.0	57.0	76.3	90.5	96.4	99.3	99.7	100.0	100.0	21.6	32.4	54.4	79.2	94.2	98.9	99.9	100.0	100.0	100.0	100.0

One-sided probitnormal Wald test

n	250											500										
$\sigma \mid \mu$	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5	-0.5	-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5
0.75	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3
0.80	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	1.7	4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	1.0	5.0
0.85	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.6	2.7	6.9	15.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	1.5	7.2	25.6
0.90	0.0	0.0	0.0	0.0	0.0	0.4	1.2	3.4	8.6	19.3	33.5	0.0	0.0	0.0	0.0	0.0	0.1	0.6	2.3	10.0	30.3	59.9
0.95	0.0	0.0	0.0	0.1	0.7	1.9	5.1	11.6	22.5	37.7	59.1	0.0	0.0	0.0	0.0	0.2	1.0	3.8	13.5	35.4	64.4	88.0
1.00	0.0	0.1	0.3	1.0	2.8	6.4	13.8	25.5	40.8	60.6	78.8	0.0	0.0	0.0	0.3	1.4	5.9	17.1	40.8	67.3	88.6	97.7
1.05	0.1	0.8	1.6	3.7	8.4	16.2	28.1	43.1	62.6	79.0	92.0	0.0	0.1	0.5	2.3	8.4	20.6	44.4	70.2	89.6	98.0	99.6
1.10	1.0	2.2	4.5	10.3	18.8	30.3	45.8	64.3	80.1	92.0	97.1	0.2	0.9	3.4	9.5	24.6	48.5	73.4	90.2	97.9	99.6	100.0
1.15	3.1	6.2	12.1	20.7	32.5	47.9	65.5	80.0	92.0	96.9	98.8	1.6	4.5	12.8	28.8	52.5	75.7	90.8	98.1	99.7	100.0	100.0
1.20	8.2	14.2	22.9	34.0	50.9	66.6	80.9	92.6	96.8	98.7	99.6	6.1	15.9	33.0	55.9	77.2	91.3	98.2	99.7	99.9	100.0	100.0
1.25	16.2	26.3	36.5	52.9	67.7	81.9	92.3	96.7	98.7	99.5	99.9	18.8	36.5	58.4	79.1	92.0	98.3	99.7	99.9	100.0	100.0	100.0

Table 2: Estimated size and power of one-sided probitnormal score test (top row) and Wald test (bottom row) at varying n. Results are in percentage and are obtained using 1000 simulations.

3.6 Moment test as a special case of the bispectral test

Dowd (2008) proposed extending the Berkowitz (2001) test to further test the skewness and kurtosis parameter, where they claim that such a test has more power in detecting different forms of forecast distribution misspecification.

Similarly, for the spectral test in Section 3.4.3, we can further devise a test for both the mean and variance of $W_{v,t}$, with the hypothesis

$$\begin{aligned} H_0: &\quad \mathrm{E}\left(W_{v,t}\right) = \mu_v \quad \text{and} \quad \mathrm{var}(W_{v,t}) = \sigma_v^2, \\ H_1: &\quad \mathrm{E}\left(W_{v,t}\right) \neq \mu_v \quad \text{or} \quad \mathrm{var}(W_{v,t}) \neq \sigma_v^2. \end{aligned} \tag{3.97}$$

We denote $Y_{v,t} = (W_{v,t} - \mu_v)^2$ and $\hat{s}_v = \frac{1}{n}\sum_{t=1}^{n} Y_{v,t}$. The test statistic is given by

$$S = \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \Sigma^{-1} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right), \tag{3.98}$$

which is asymptotically χ_2^2 distributed, where

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\mu}_v - \mu_v \\ \hat{s}_v \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} 0 \\ \sigma_v^2 \end{pmatrix}, \tag{3.99}$$

and

$$\Sigma = \frac{1}{n} \begin{pmatrix} \sigma_v^2 & \mathrm{E}\left((W_{v,t} - \mu_v)^3\right) \\ \mathrm{E}\left((W_{v,t} - \mu_v)^3\right) & \mathrm{E}\left((W_{v,t} - \mu_v)^4\right) - \sigma_v^4 \end{pmatrix}. \tag{3.100}$$

For the continuous weighting case, $\mathrm{E}\left((W_{v,t} - \mu_v)^3\right)$ and $\mathrm{E}\left((W_{v,t} - \mu_v)^4\right)$ can be calculated using Lemma 3.4. We can generalize the above test to test for all higher moments.

Note that since $Y_{v,t} = W_{v,t}^2 - 2\mu_v W_{v,t} + \mu_v^2$, for the continuous weighting case, we know that $W_{v,t}^2 = W_{g_2^*,t}$ where g_2^* is given in Proposition 3.3. Hence, the above test is same as testing linear combinations of $W_{v,t}$ and $W_{g_2^*,t}$.

3.7 How the Pearson chi-squared test relates to the k-spectral test

For a series $(P_1, \ldots, P_n)$, we denote $W_{i,t} = \int_0^1 I_{\{P_t > u\}} \mathrm{d}\nu_i(u)$, where $\nu_i = \delta_{\alpha_i} - \delta_{\alpha_{i+1}}$, for $i = 0, 1, \ldots, N$, where $\alpha_0 < \alpha_1 < \ldots < \alpha_{N+1}$ are ordered levels with $\alpha_0 = 0$ and $\alpha_{N+1} = 1$.

$(W_{i,1}, \ldots, W_{i,n})$ form a series of iid Bernoulli(p_i) variables, with $p_i = \alpha_{i+1} - \alpha_i$. We denote by $Z_i = \frac{\sum_{t=1}^{n}(W_{i,t}-p_i)}{\sqrt{n\,p_i\,(1-p_i)}}$ the normalized sum of $W_{i,t}$. By the central limit theorem, for $i = 1, \ldots N$, the Z_i are standard normal variables, so that $\tilde{N}_i = \sqrt{1-p_i}\,Z_i$ are $N(0, 1-p_i)$ distributed. It can be shown that $\mathrm{cov}(\tilde{N}_i, \tilde{N}_j) = -\sqrt{p_i\,p_j}$ for $i \neq j$.

The Pearson chi-squared test is a two-sided test for the vector $\boldsymbol{v} = (\tilde{N}_0, \tilde{N}_1, \ldots, \tilde{N}_N)^T$. The covariance matrix of $\boldsymbol{v}$ has rank N, i.e. it is not invertible. Pearson (1900) has shown that by using a new coordinate system with the vector $\boldsymbol{p} = (\sqrt{p_0}, \sqrt{p_1}, \ldots, \sqrt{p_N})^T$ as the first basis vector, the new coordinate vector is $\tilde{\boldsymbol{v}} = (0, \tilde{Z}_1, \ldots, \tilde{Z}_N)^T$, where $(\tilde{Z}_1, \ldots, \tilde{Z}_N)$ are iid standard normal variables. Hence, the chi-squared test statistic in (3.17) is $S_N = \tilde{\boldsymbol{v}}^T\tilde{\boldsymbol{v}} = \sum_{i=1}^{N}\tilde{Z}_i^2$, which is asymptotically χ^2_N distributed.

Chapter 4 Simulation studies: Static tests

In this chapter, we will attempt to understand the characteristics of the static tests described in Chapter 3 via simulations. To facilitate analysis, the following colour coding is used in most of the tables: green indicates good results ($\leqslant 6\%$ for the size; $\geqslant 70\%$ for the power); red indicates poor results ($\geqslant 9\%$ for the size; $\leqslant 30\%$ for the power); dark red indicates very poor results ($\geqslant 12\%$ for the size; $\leqslant 10\%$ for the power).

4.1 In the case when there is no parameter estimation error

In this section, we will attempt to understand the characteristics of the static tests described in Chapter 3 in the ideal situation when no parameter estimation error is involved.

4.1.1 Experimental design

In each experiment we generate a total dataset of n values from the true distribution F_t; we consider the cases when F_t is normal, Student-t distributions with five and three degrees of freedom, denoted by $t5$ and $t3$, which have moderately heavy and heavy tails respectively, and the skewed Student-t distribution of Fernandez & Steel (1998) with three degrees of freedom and a skewness parameter $\gamma = 1.2$, denoted s$t3$. We standardized F_t to have zero mean and unit variance. The forecast distribution $\widehat{F}_t$ is always the standard normal distribution.

4.1.2 Binomial test results

We first look at the size and power when we apply the one-sided and two-sided binomial score test and LRT as described in Section 3.2.1 to the exceptions of VaR estimates at level $\alpha = 0.975$ and $\alpha = 0.99$. For the score test, we denote by W_{θ_0} when the true Fisher information $I(\theta_0)$ is used, with $\theta_0 = 1 - \alpha$, and $W_{\hat{\theta}}$ when the estimated Fisher information $I(\hat{\theta})$ is used, where $\hat{\theta}$ is the maximum likelihood

estimator for the binomial distribution.

Table 3 shows the values of $\text{VaR}_{0.975}$ and $\text{VaR}_{0.99}$ for the four distributions used in the simulation study. These distributions have all been standardized to have mean zero and variance one. Δ shows the percentage increase in the value of $\text{VaR}_{0.99}$ when compared with the normal distribution. Table 4 shows the results for one-sided and two-sided binomial score tests at the 97.5% and 99% levels.

	$\text{VaR}_{0.975}$	$\text{VaR}_{0.99}$	Δ
Normal	1.96	2.33	0.00
$t5$	1.99	2.61	12.04
$t3$	1.84	2.62	12.69
s$t3$ ($\gamma = 1.2$)	2.04	2.99	28.68

Table 3: Values of $\text{VaR}_{0.975}$ *and* $\text{VaR}_{0.99}$ *for four distributions used in simulation study (normal, Student-t5, Student-t3, skewed Student-t3 with skewness parameter* $\gamma = 1.2$*).* Δ *column shows percentage increase in* $\text{VaR}_{0.99}$ *compared with normal distribution.*

97.5% level. The size of the tests is generally reasonable. W_{θ_0} in particular always seems to have a good size for all the different sample sizes in both the one-sided and two-sided tests. For the score test, using $I(\hat{\theta})$ instead of $I(\theta_0)$ reduces the speed of convergence of the Z-test statistic, leading to tests that are either undersized or oversized when sample size n is small. The power of all the tests is very weak, since the 97.5% VaR values of all the distributions are quite similar.

99% level. At this level the size is usually too large in the smaller samples. The binomial score test W_{θ_0} seems to have the best size performance.

At this level, the tests are more powerful because the differences between the quantiles of the four models are larger. One-sided tests are somewhat more powerful than two-sided tests. The score test and LRT seem to be more powerful than the Wald test.

	α	0.975						0.990					
	twosided	TRUE			FALSE			TRUE			FALSE		
F_t	n | test	$W_{\hat{\theta}}$	W_{θ_0}	LRT	$W_{\hat{\theta}}$	W_{θ_0}	LRT	$W_{\hat{\theta}}$	W_{θ_0}	LRT	$W_{\hat{\theta}}$	W_{θ_0}	LRT
Normal	250	5.7	3.9	7.5	2.4	5.0	5.0	8.0	4.0	8.9	1.2	4.0	4.0
	500	7.8	3.9	5.9	2.6	4.7	4.7	12.5	3.7	7.0	1.3	6.7	3.1
	1000	5.0	5.0	4.1	2.8	4.3	4.3	7.5	3.8	5.9	2.7	4.9	4.9
	2000	5.9	5.0	4.2	3.9	5.0	5.0	4.9	5.4	4.1	3.5	5.3	5.3
t5	250	4.3	4.1	6.9	3.1	6.4	6.4	5.9	17.7	10.7	8.3	17.7	17.7
	500	6.0	5.2	6.5	4.4	7.4	7.4	9.5	22.4	22.8	13.4	33.9	22.3
	1000	4.9	6.9	5.2	5.7	8.0	8.0	17.7	33.0	33.1	33.0	42.7	42.7
	2000	6.0	7.3	5.8	8.3	10.7	10.7	45.3	59.9	52.7	59.9	66.7	66.7
t3	250	9.6	3.6	10.3	0.8	2.0	2.0	5.6	13.5	9.2	6.0	13.5	13.5
	500	15.8	4.8	9.5	0.6	1.3	1.3	7.8	16.2	16.9	9.3	25.4	16.1
	1000	14.2	9.9	9.7	0.4	0.6	0.6	11.0	22.3	22.5	22.2	30.5	30.5
	2000	25.9	16.6	16.5	0.2	0.3	0.3	27.6	41.4	34.2	41.3	48.8	48.8
st3	250	4.4	5.4	8.0	4.5	8.6	8.6	10.4	31.2	19.2	18.3	31.2	31.2
	500	6.0	6.9	7.9	6.3	10.1	10.1	22.4	44.2	44.3	31.9	57.2	44.2
	1000	5.5	9.5	6.9	9.0	12.3	12.3	48.6	66.2	66.2	66.2	74.7	74.7
	2000	8.4	12.2	9.8	14.6	17.9	17.9	86.6	92.9	90.1	92.9	95.0	95.0

Table 4: Estimated size and power of three different types of binomial test applied to exceptions of the 97.5% and 99% VaR estimates. Both one-sided and two-sided tests have been carried out. Results are based on 10000 replications

4.1.3 Theoretical rejection rate of the one-sided Binomial score test

We will attempt to understand the results in Table 4 better. The finite sample rejection rate of the one-sided binomial score test at level α with approximately size κ is given by

$$\begin{aligned} \mathrm{P}\left(Z_{\text{score}} > \Phi^{-1}(1-\kappa)\right) &= P\left(\sum_{t=1}^{n} C_t > \sqrt{n\alpha(1-\alpha)}\Phi^{-1}(1-\kappa) + n(1-\alpha)\right) \\ &= 1 - B_{n,1-\theta_\alpha}\left(\sqrt{n\alpha(1-\alpha)}\Phi^{-1}(1-\kappa) + n(1-\alpha)\right) \quad (4.1) \end{aligned}$$

with C_t defined in (3.8), Z_{score} is given in (3.12), and $B_{n,1-\theta_\alpha}$ is the binomial distribution function with size parameter n and success probability $1-\theta_\alpha$. We have assumed that $(C_1, \ldots, C_n)$ are iid Bernoulli$(1-\theta_\alpha)$ variables, where

$$\theta_\alpha = F_t\left(\widehat{F}_t^{-1}(\alpha)\right), \quad (4.2)$$

$\widehat{F}_t$ and F_t are the forecast and true distribution functions as described in Section 4.1.1. The exception rate at level α is given by

$$\mathrm{P}\left(P_t > \alpha\right) = \mathrm{P}\left(\widehat{F}_t(F_t^{-1}(U)) > \alpha\right) = 1 - \theta_\alpha, \quad (4.3)$$

where U is a standard uniform random variable.

Figure 1 plots the exception rate $1-\theta_\alpha$ and the binomial score test rejection rate, when $\widehat{F}_t$ is normal, and F_t is normal, Student-$t5$, Student-$t3$, skewness Student-$t3$ with skewness parameter $\gamma = 1.2$, with α in the interval $(0.9, 1)$, and $n = 1000$.

The result in Figure 1 is consistent to the simulated results in Table 4. Also, notice that at very high level, the score test becomes increasingly over-sized. To understand the rate of convergence of Z_{score} to the standard normal distribution, we use the Berry-Esseen theorem, where for all x and n,

$$|\widehat{F}_n^\alpha(x) - \Phi(x)| \leqslant \frac{C}{\sqrt{n}} \frac{\alpha(1-\alpha)(1-2\alpha(1-\alpha))}{(\alpha(1-\alpha))^{3/2}} = \frac{C}{\sqrt{n}} R_\alpha, \quad (4.4)$$

where $\widehat{F}_n^\alpha(x)$ denotes the empirical cdf of Z_{score} at level α, $\Phi(x)$ is the cdf of a standard normal distribution, and C is some constant. Figure 2 shows the ratio R_α for α ranging between 0.9 to 1. We see that in order for Z_{score} at level $\alpha = 0.975$ to have the same convergence rate with Z_{score} at level $\alpha = 0.9$, we need roughly $2.23^2 = 4.97$ times more data. Similarly, for Z_{score} at level $\alpha = 0.99$ to have the same convergence rate with Z_{score} at level $\alpha = 0.9$, we need roughly $3.60^2 = 12.96$ times more data.

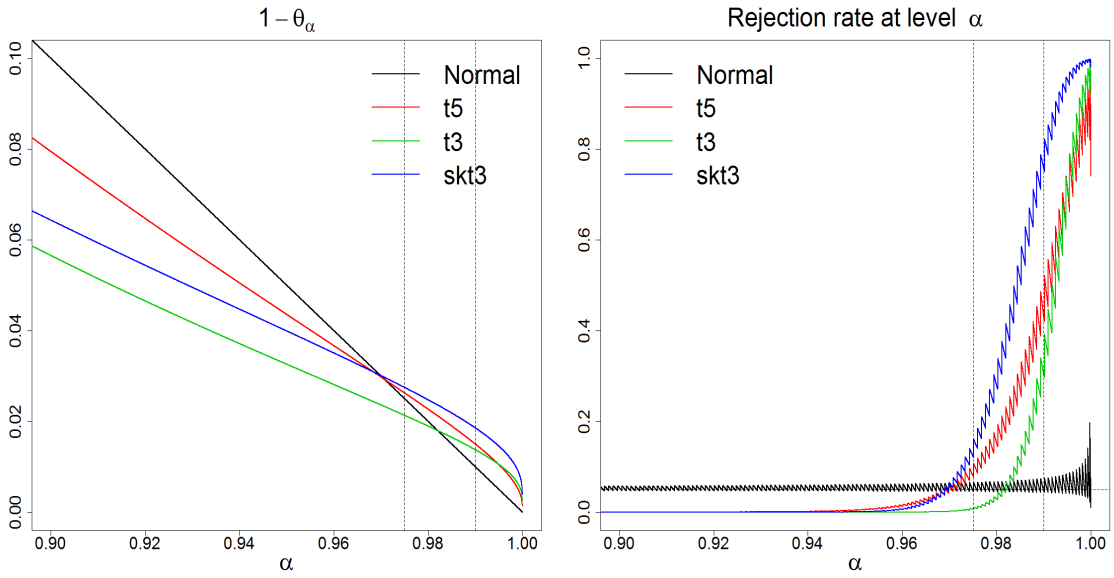

Figure 1: The exception rate $1 - \theta_\alpha$ *and the binomial score test rejection rate, when* $\widehat{F}_t$ *is normal, and* F_t *is normal, Student-t5, Student-t3, skewed Student-t3 with skewness parameter* $\gamma = 1.2$*, with* α *in the interval* $(0.9, 1.0)$*, and* $n = 1000$*. The dotted vertical lines represents the levels* $\alpha = 0.975$ *and* $\alpha = 0.99$*.*

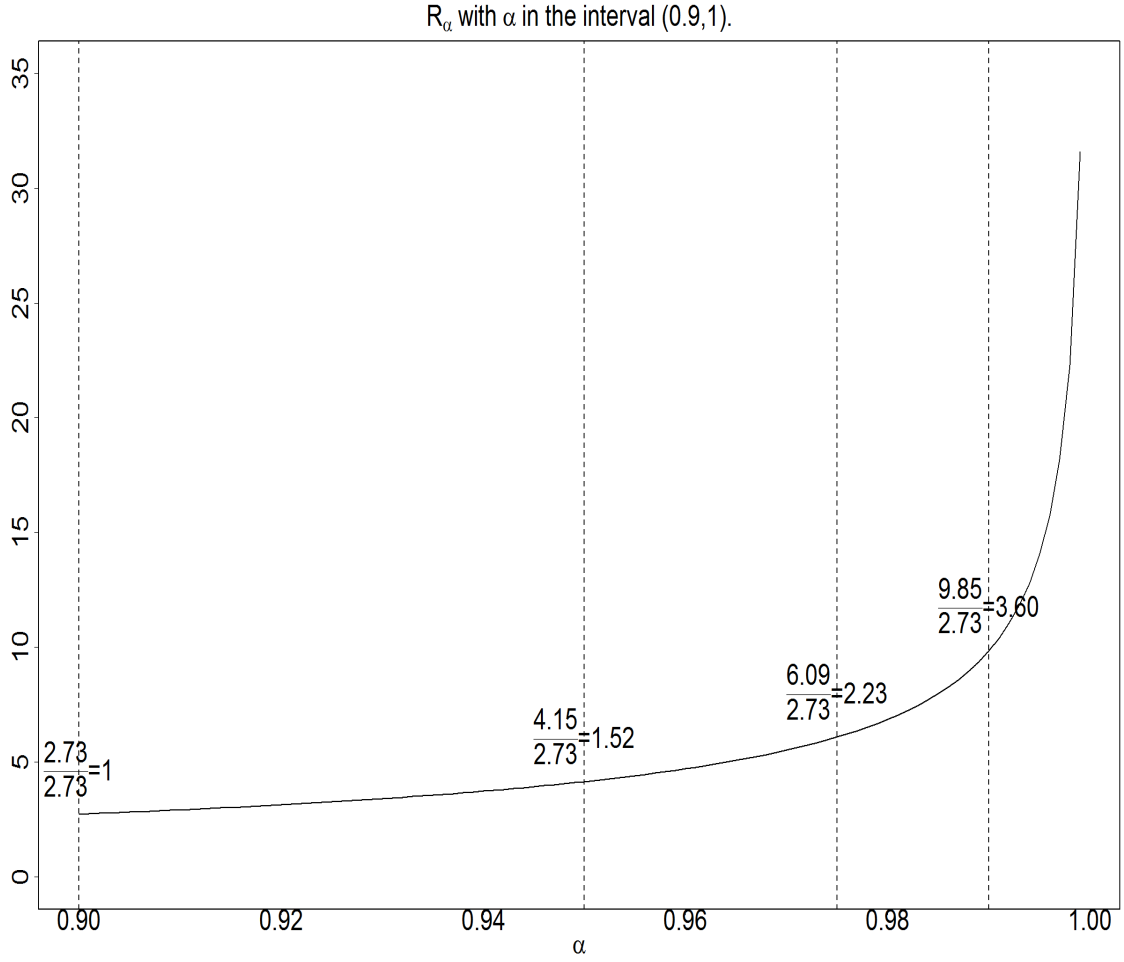

Figure 2: The ratio R_α *with* α *in the interval* $(0.9, 1.0)$*.*

4.1.4 Multinomial test

From Section 4.1.3, we observed that although the binomial test is more powerful when the level α gets close to one, the size performance becomes worse at the same time. By testing VaR exception rate at multiple levels using the multinomial tests, we hope to achieve some balance between power and size. To determine the exception rate level values we set $N = 2^k$ for $k = 0, 1, \cdots, 6$. In all multinomial experiments with $N \geqslant 2$ we set $\alpha_1 = \alpha = 0.975$ and further levels are determined by

$$\alpha_j = \alpha + \frac{j-1}{N}(1-\alpha)\,, \quad j = 1, \ldots, N, \;\; N \in \mathbb{N}\,. \tag{4.5}$$

The choice $\alpha = 0.975$ is motivated by the regulatory requirement of testing $\mathrm{ES}_{0.975}$, and the multinomial test can be seen as an implicit test for the expected shortfall. See Kratz et al. (2016). We choose sample sizes $n = 250, 500, 1000, 2000$ and estimate the rejection probability for the null hypothesis using 10,000 replications.

The results for the multinomial tests described in Section 3.2.2 are shown in Table 5. As discussed in Section 3.2.2, in the case $N = 1$, the Pearson test gives identical results to the two-sided score test W_{θ_0} in Table 4, and the Nass statistic is very close to the value of the Pearson statistic (the scaling constant c in (3.18) is slightly less than one) and also gives much the same results. The LRT with $N = 1$ is the two-sided LRT from Table 4.

	test	Pearson							Nass							LRT						
F_t	$n \mid N$	1	2	4	8	16	32	64	1	2	4	8	16	32	64	1	2	4	8	16	32	64
Normal	250	3.9	4.7	5.6	8.5	10.5	14.1	21.5	3.9	3.5	5.0	4.7	5.1	5.0	4.8	7.5	10.0	6.5	6.5	6.5	6.2	6.1
	500	3.9	4.4	5.2	6.6	8.6	12.3	16.2	3.9	3.9	4.7	4.7	5.5	5.5	5.3	5.9	5.8	5.5	5.6	5.3	5.3	5.2
	1000	5.0	5.2	5.0	5.6	7.2	9.0	12.0	5.0	4.8	4.7	4.9	5.1	5.3	5.1	4.1	5.5	5.5	5.8	5.6	5.6	5.7
	2000	5.0	4.5	4.8	5.0	6.3	7.2	8.8	5.0	4.3	4.5	4.5	5.3	5.1	4.9	4.2	4.9	4.7	5.0	5.1	5.1	5.0
$t5$	250	4.1	10.2	14.1	20.8	22.4	27.0	34.2	4.1	7.7	12.8	14.1	13.4	14.4	13.0	6.9	14.4	15.8	21.6	26.6	30.7	33.7
	500	5.2	15.7	22.1	28.4	32.2	36.2	39.8	5.2	14.3	20.5	24.5	26.6	26.0	22.7	6.5	15.5	26.9	36.6	44.7	50.4	54.8
	1000	6.9	26.7	40.2	48.2	53.0	54.8	55.8	6.9	25.5	39.5	46.2	48.6	47.7	43.8	5.2	26.1	46.4	61.8	71.4	76.7	80.5
	2000	7.3	47.2	70.4	79.3	82.5	82.8	82.0	7.3	47.0	69.6	78.2	80.8	80.2	77.0	5.8	48.0	77.4	89.5	94.4	96.6	97.6
$t3$	250	3.6	7.3	13.7	21.1	19.4	25.8	28.1	3.6	5.6	12.1	14.8	13.4	13.2	13.6	10.3	24.4	24.4	35.4	43.2	48.0	51.9
	500	4.8	16.1	25.2	32.7	35.2	40.1	38.6	4.8	15.5	22.4	28.7	32.3	29.4	26.4	9.5	26.2	44.2	58.6	67.9	73.8	78.0
	1000	9.9	37.4	55.6	62.9	65.2	64.8	64.2	9.9	35.2	54.1	60.3	61.4	59.9	54.7	9.7	47.2	75.4	87.7	93.2	95.5	96.8
	2000	16.6	73.1	91.0	94.5	94.9	93.9	92.1	16.6	72.7	90.5	94.2	94.3	92.6	89.6	16.5	79.5	96.8	99.4	99.8	99.9	100.0
$st3$	250	5.4	18.9	28.8	40.0	38.7	46.3	50.5	5.4	15.3	26.3	30.5	30.2	30.5	30.7	8.0	24.6	33.5	46.5	55.1	60.8	65.4
	500	6.9	34.9	50.7	60.6	64.6	69.5	70.2	6.9	33.2	47.6	56.2	61.4	60.0	56.8	7.9	35.9	59.3	73.6	81.6	86.2	88.9
	1000	9.5	62.3	83.0	89.1	91.3	92.1	92.0	9.5	61.4	82.3	88.1	90.0	90.0	87.9	6.9	62.3	88.1	95.3	97.9	98.9	99.2
	2000	12.2	90.7	98.7	99.7	99.8	99.8	99.7	12.2	90.7	98.6	99.7	99.7	99.7	99.5	9.8	91.6	99.3	99.9	100.0	100.0	100.0

Table 5: Estimated size and power of three different types of multinomial test (Pearson, Nass, likelihood-ratio test (LRT)) based on exceptions of N levels. Results are based on 10000 replications

Size of the tests. The results for the size of the three tests are summarized in the first panel of Table 5 where F_t is normal. We observed that the size of the Pearson χ^2-test is poor for large number of levels ($N \geqslant 8$). As we would expect, the Nass test, which is the size-corrected version of the Pearson test, has the best size properties, where the sizes are very stable for all choices of N and all sample sizes. The LRT tends to be over-sized in smaller sample size ($n = 250$) but otherwise has reasonable size performance for all choices of N.

Power of the tests. The results for the power of the three tests are summarized in the panels 2–4 of Table 5. It can be seen that for all N the LRT is generally the most powerful test. The Nass test has similar power to the Pearson test (at $N \leqslant 4$ when the Pearson test has acceptable size). Also, as we would have expect, as the tail of F_t gets fatter, the tests become more powerful.

It seems clear that, regardless of the test chosen, multinomial tests with $N \geqslant 2$ are much more powerful than a binomial test. Another advantage of the multinomial test over the binomial test is that, as we have seen in Figure 1, the results from binomial tests are much more sensitive to the choice of α. By using a range of levels the multinomial tests are much less sensitive to the exact choice of these levels, which makes them a more reliable type of tests.

4.1.5 Spectral and bispectral test with different weight functions

In the previous section, we have seen the benefits of testing VaR exceptions at multiple levels. In this section, we explore the size and power of the spectral and bispectral tests as described in Section 3.4.3 and Section 3.4.4, which test the weighted integral of VaR exceptions in the range $[\alpha_1, \alpha_2]$.

The structure of the experiment is the same as before, where we assume $\widehat{F}_t$ to be standard normal, and F_t to be normal, Student-$t5$, Student-$t3$, skewed Student-$t3$ with skewness parameter $\gamma = 1.2$, standardized to have zero mean and unit variance, with $n = 250, 500, 1000, 2000$. We set $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$. There are three reasons which motivates the choice of setting $\alpha_2 = 0.9995$. The first reason is

that it is very difficult to calibrate the forecast distribution $\widehat{F}_t$ to model extreme tails accurately in practice, due to data availability. The second reason is that in the event that the realized loss exceeds some extreme level, it is very unlikely that the financial institution will survive the loss anyway. Finally, when empirical methods such as the historical simulation model is used as the forecast distribution, we obtain $P_t = 1$ when the realized loss exceeds the largest value of the data used for calibration, which may lead to the test statistic of some tests (for example, the Berkowitz (2001) test) being undefined.

We consider the following tests for the simulation study:

Binomial score test at 99% level, denoted by B99.

Spectral tests at uniform, linear and exponential weight (with $k = 200$), denoted by SP.U, SP.L and SP.E.200 respectively. See Section 3.4.3 for more details. Figure 3 plots the weight functions, rescaled to unit area.

Bispectral tests at uniform-linear, uniform-exponential (with $k = 200$), and linear-exponential (with $k = 100, 200$) weight functions, denoted by SP.UL, SP.UE.200, SP.LE.100 and SP.LE.200 respectively. See Section 3.4.4 for more details.

truncated probitnormal score test as described in Section 3.5, denoted by PNS.

Berkowitz tests as proposed in Berkowitz (2001), truncated to the range (0.975,0.9995), denoted by BK.

The size and power of the tests are shown in Table 7. To summarize:

Size of the tests. The results for the size of the tests are summarized in the first panel of Table 7 where F_t is normal. We observe that all tests have acceptable size. For small sample size ($n = 250$), tests that place more weight towards the tail become increasingly oversized.

Power of the tests. The results for the power of the tests are summarized in the panels 2–4 of Table 7. We observed that the binomial score test is more powerful

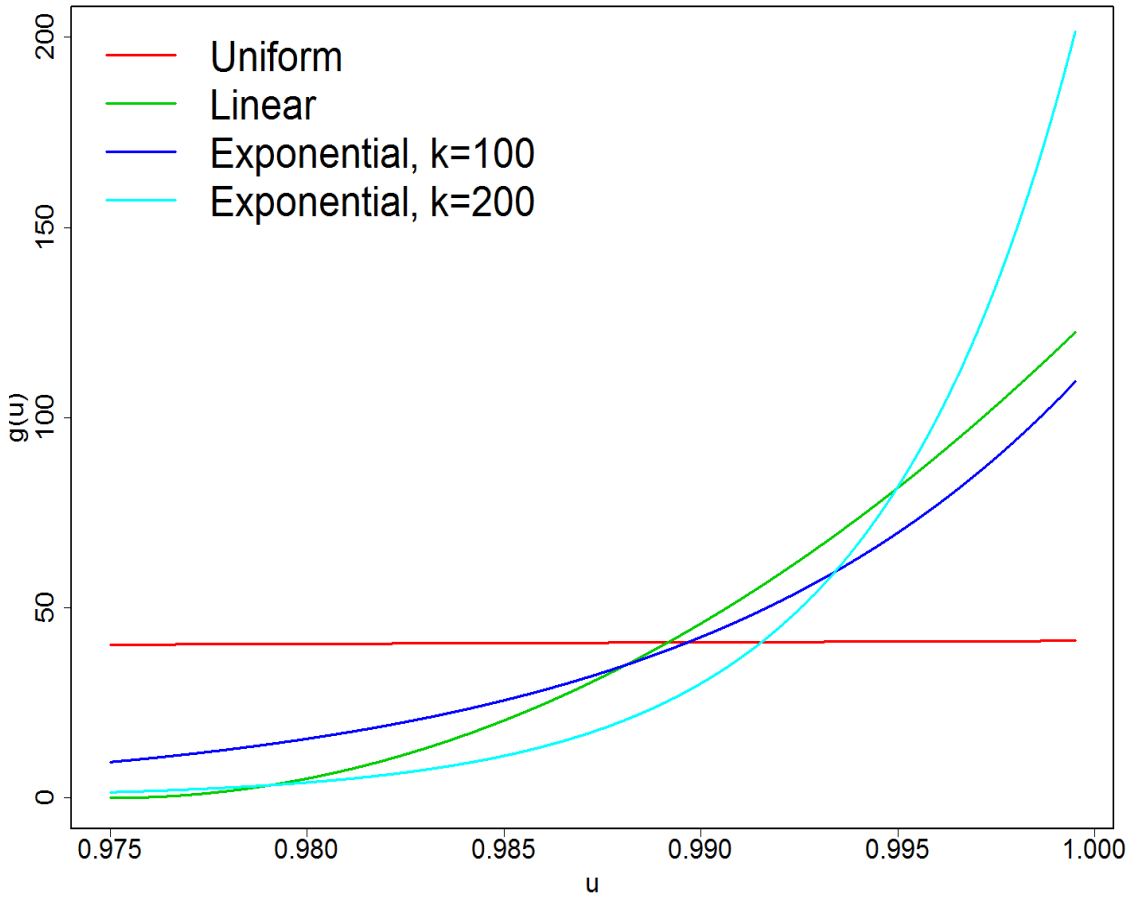

Figure 3: The above weight functions, rescaled to unit area, for $\alpha_1 = 0.975$ *and* $\alpha_2 = 0.9995$. *We set* $k = 200$ *for the exponential weight.*

than the uniform spectral test. However, spectral tests that place more weight towards the tail become increasingly powerful.

Another observation is that the bispectral tests are a lot more powerful than the spectral tests. We speculate that this is due to the extra information gained from the correlation between $W_{v_1,t}$ and $W_{v_2,t}$, which we denote by ρ_{v_1,v_2}. This observation is supported by two points:

- Despite SP.E.200 placing more weight towards the tail on average compared to SP.UL, SP.UL is more powerful than SP.E.200.

- Despite SP.LE.100 placing more weight towards the tail on average compared to SP.UL, there are cases when SP.UL is more powerful. This is due to the fact that the linear function and exponential function at $k = 100$ are quite similar (See Figure 3), hence the correlation is close to one, i.e. the information gain from this combination of weight function is small.

Table 6 shows the correlation ρ_{v_1,v_2}, and average proportion of weight placed in different regions, for the various combinations of weight functions. $w_{(x,y)}$ represents the average proportion of weight placed in the range (x,y), obtained using the equation $\int_x^y \big(\tilde{g}_1(u) + \tilde{g}_2(u)\big)/2 \, \mathrm{d}u$, and w_x represents the average proportion of weight placed at the point x, obtained using the equation $\big(\tilde{v}_1(x) + \tilde{v}_2(x)\big)/2$, where $\tilde{g}_i$ and $\tilde{v}_i$ are the normalized weight function and weight measure such that $\int_0^1 \tilde{v}_i(u) \, \mathrm{d}u = 1$.

test	SP.UL	SP.UE.200	SP.LE.100	SP.LE.200	PNS
ρ_{v_1,v_2}	96.8	87.2	99.9	95.7	97.5
$w_{0.975}$	0.0	0.0	0.0	0.0	40.5
$w_{(0.975,0.99)}$	49.3	37.6	35.1	25.9	11.7
$w_{(0.99,0.995)}$	24.8	23.2	28.4	27.5	8.5
$w_{(0.995,0.9995)}$	25.9	39.2	36.5	46.6	27.7
$w_{0.9995}$	0.0	0.0	0.0	0.0	11.6

Table 6: The correlation between $W_{v_1,t}$ and $W_{v_2,t}$, denoted by ρ_{v_1,v_2}, and the average proportion of weight placed in different regions, for the various combinations of weight functions, when $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$. Units in percentage.

Both the PNS and BK have good power performance, with PNS being slightly more powerful. Their performance is similar, which is to be expected since as we have explained in Section 3.5.1, BK is derived from the two sided probitnormal score test

with some additional approximations.

The discrete probitnormal LRT can be viewed as a discretized version of the BK test. We denote by I_{DBK} and I_{BK} the rejection rate indicators of the discrete probitnormal LRT and BK, which takes the value one when the test rejects the null hypothesis for a particular simulated sequence of (P_t). The rejection rate (and hence size and power) of the discrete probitnormal LRT with levels as defined in Section 4.1.4 is the same as the BK test truncated to the range (α_1,α_N). Hence, in the case when $N = 50$, the discrete probitnormal LRT has the same size and power as the BK test truncated to the range (0.975,0.9995), i.e. $\mathrm{E}\,(I_{\mathrm{DBK}}) = \mathrm{E}\,(I_{\mathrm{BK}})$. In this case, they differ in the variance of the rejection rate indicator, with $\mathrm{var}(I_{\mathrm{DBK}}) > \mathrm{var}(I_{\mathrm{BK}})$. In other words, for a given sequence of (P_t), assuming that the range and levels are set so that the BK test and discrete probitnormal LRT have equal size and power, the result of the BK test is more reliable than those of the discrete probitnormal LRT.

F_t	n | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	SP.LE.100	SP.LE.200	PNS	BK
Normal	250	4.0	4.4	4.0	4.2	4.9	5.1	5.1	5.3	5.6	6.2
	500	3.7	4.6	4.7	4.5	4.9	4.9	4.8	5.0	5.0	5.3
	1000	3.8	5.3	5.3	5.1	5.1	4.9	5.2	4.8	5.1	5.7
	2000	5.4	4.8	4.8	4.9	5.1	4.8	5.0	5.0	4.7	5.0
$t5$	250	17.7	16.0	23.6	33.8	30.1	36.9	32.6	38.4	42.6	32.2
	500	22.4	21.1	33.3	47.9	45.5	54.2	45.5	55.7	62.1	53.1
	1000	33.0	30.2	49.5	69.0	69.6	77.5	66.1	78.6	83.9	79.4
	2000	59.9	48.5	73.3	90.2	92.7	96.0	88.2	96.1	98.0	97.3
$t3$	250	13.5	11.2	20.3	35.2	37.6	46.4	35.3	48.0	54.3	50.8
	500	16.2	12.6	27.6	48.7	61.4	69.7	50.1	69.8	78.5	76.7
	1000	22.3	15.9	39.7	69.0	88.6	92.6	72.1	92.0	96.7	96.3
	2000	41.4	21.3	59.7	89.5	99.5	99.8	92.1	99.7	100.0	100.0
$st3$	250	31.2	27.3	41.8	57.8	55.7	64.5	56.3	66.2	71.8	63.8
	500	44.2	39.8	60.5	77.8	80.1	86.6	77.2	87.1	91.2	88.2
	1000	66.2	60.6	83.3	95.2	97.1	98.5	94.5	98.5	99.4	99.1
	2000	92.9	85.0	97.5	99.8	100.0	100.0	99.8	100.0	100.0	100.0

Table 7: Estimated size and power of various two-sided tests, based on 10000 Replications, with $\alpha_1 = 0.975$ *and* $\alpha_2 = 0.9995$.

4.1.6 One-sided spectral and bispectral tests

In this section we consider the corresponding one-sided versions of the spectral and bispectral tests which we have analyzed in the previous section. The construction of these tests is described in Section 3.4.5. We have also included the one-sided probitnormal Wald test based on Section 3.5.2, which we denote by PN.Wald, and the one-sided binomial score test for comparison. The size and power of these tests are shown in Table 8.

Size of the tests. The results for the size of the tests are summarized in the first panel of Table 8 where F_t is normal. We observe that the size of both one-sided spectral and bispectral tests are worse compared to their two-sided counterparts in Table 7. PN.Wald has better size performance compared to PNS, as we would expect based on the analysis in Section 3.5.2 (the result is similar to those in Table 2 when $\mu = 0$ and $\sigma = 1$).

Power of the tests. The results for the power of the tests are summarized in the panels 2–4 of Table 8. We observe that the one-sided tests are more powerful than their two-sided counterparts. This is because the forecast model that we consider (the standard normal distribution) has higher probability of underestimating the region (the range [0.975, 0.9995]) of the distributions that we are testing. As we would expect, PNS has better power performance compared to PN.Wald.

F_t	n | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	SP.LE.100	SP.LE.200	PNS	PN.Wald
Normal	250	4.0	6.1	6.3	6.6	5.7	6.5	6.5	6.4	7.3	5.4
	500	6.7	5.8	6.1	6.7	5.6	6.2	6.4	6.3	6.5	5.8
	1000	4.9	5.8	6.0	6.3	5.3	5.7	5.8	5.8	5.8	6.1
	2000	5.3	5.3	5.6	5.6	5.2	5.5	5.3	5.3	5.6	5.5
$t5$	250	17.7	21.6	30.1	40.6	36.4	43.4	37.0	44.3	49.1	44.7
	500	33.9	28.3	41.1	55.6	53.2	61.7	50.8	61.8	68.6	60.2
	1000	42.7	39.5	58.8	76.0	76.3	82.9	70.5	82.4	88.6	78.6
	2000	66.7	58.5	80.4	93.4	95.2	97.4	90.1	97.1	98.9	93.9
$t3$	250	13.5	14.8	26.3	42.2	46.8	54.3	40.6	54.9	60.5	47.9
	500	25.4	17.5	35.0	56.1	70.6	76.9	56.1	76.5	82.1	60.2
	1000	30.5	21.7	48.4	75.5	93.1	95.1	77.0	94.8	96.9	76.2
	2000	48.8	28.9	68.2	92.9	99.8	99.9	94.4	99.8	99.9	90.5
$st3$	250	31.2	34.5	49.4	64.1	62.8	70.7	60.4	71.3	76.5	69.6
	500	57.2	48.9	68.1	82.8	85.4	90.1	80.2	89.7	93.8	86.0
	1000	74.7	68.8	88.1	96.7	98.2	99.1	95.2	98.9	99.7	97.2
	2000	95.0	90.1	98.6	99.9	100.0	100.0	99.8	100.0	100.0	99.9

Table 8: Estimated size and power of various one-sided tests, based on 10000 Replications, with $\alpha_1 = 0.975$ *and* $\alpha_2 = 0.9995$.

4.1.7 The uniform spectral test in greater detail

In this section, we study in greater detail the spectral test in the simple case when the weight function is uniform, which we will refer to as the uniform spectral test.

The size of the test depends on the rate of convergence of the distribution of Z_v in (3.46) to the standard normal distribution. We can again appeal to the Berry-Esseen theorem, where for all x and n,

$$|\widehat{F}_{Z_v,n}(x) - \Phi(x)| \leqslant \frac{C}{\sqrt{n}} \frac{\mathrm{E}\left(|W_{v,t} - \mu_v|^3\right)}{\sigma_v^3} = \frac{C}{\sqrt{n}} R_v \,, \tag{4.6}$$

where $\widehat{F}_{Z_v,n}(x)$ denotes the empirical cdf of Z_v, $\Phi(x)$ is the cdf of a standard normal distribution, and C is some constant. For the continuous weighting case when the weight function is uniform, we can show that

$$\mathrm{E}\left(|W_{v,t} - \mu_v|^3\right) = \frac{1}{4}(\mu_v^4 + (\alpha_2 - \alpha_1 - \mu_v)^4) + \alpha_1\mu_v^3 + (1-\alpha_2)(\alpha_2 - \alpha_1 - \mu_v)^3 \,, \tag{4.7}$$

where $\mu_v = \left[-\frac{1}{2}(1-u)^2\right]_{u=\alpha_1}^{\alpha_2}$.

Figure 4 plots the Berry-Esseen ratio R_v of the uniform spectral test. The left plot shows R_v when we fix $\alpha_2 = 0.975$ and vary α_1 in the interval $(0.95, 0.975)$, and the right plot shows R_v when we fix $\alpha_1 = 0.975$ and vary α_2 in the interval $(0.975, 1)$. We see that as we expand the range towards the left, the convergence rate improves, and conversely, as we expand the range towards the right, the convergence rate worsen.

Next, we look at the rejection rate of the one-sided uniform spectral test. Similar to Section 4.1.3, we can approximate the rejection rate of the one-sided uniform spectral test. First, we note that for the continuous weighting case when the weight function is uniform, we can write

$$\begin{aligned} W_{v,t} &= \int_{\alpha_1}^{\min(\alpha_2, \max(P_t, \alpha_1))} \mathrm{d}u && (4.8)\\ &= \min(\alpha_2, \max(P_t, \alpha_1)) - \alpha_1 && (4.9)\\ &= I_{\{P_t < \alpha_1\}}\alpha_1 + I_{\{P_t > \alpha_2\}}\alpha_2 + I_{\{\alpha_1 \leqslant P_t \leqslant \alpha_2\}}P_t - \alpha_1 && (4.10)\\ &= A_t - \alpha_1, && (4.11) \end{aligned}$$

where we denote $A_t = I_{\{P_t < \alpha_1\}}\alpha_1 + I_{\{P_t > \alpha_2\}}\alpha_2 + I_{\{\alpha_1 \leqslant P_t \leqslant \alpha_2\}}P_t$. Appealing to the Central Limit Theorem, for large n, $\frac{1}{n}\sum_{t=1}^{n} A_t \sim N(\mu_A, \sigma_A^2)$, with

$$\mu_A = \alpha_1\mu_L + \alpha_2\mu_R + \mu_M \,, \tag{4.12}$$

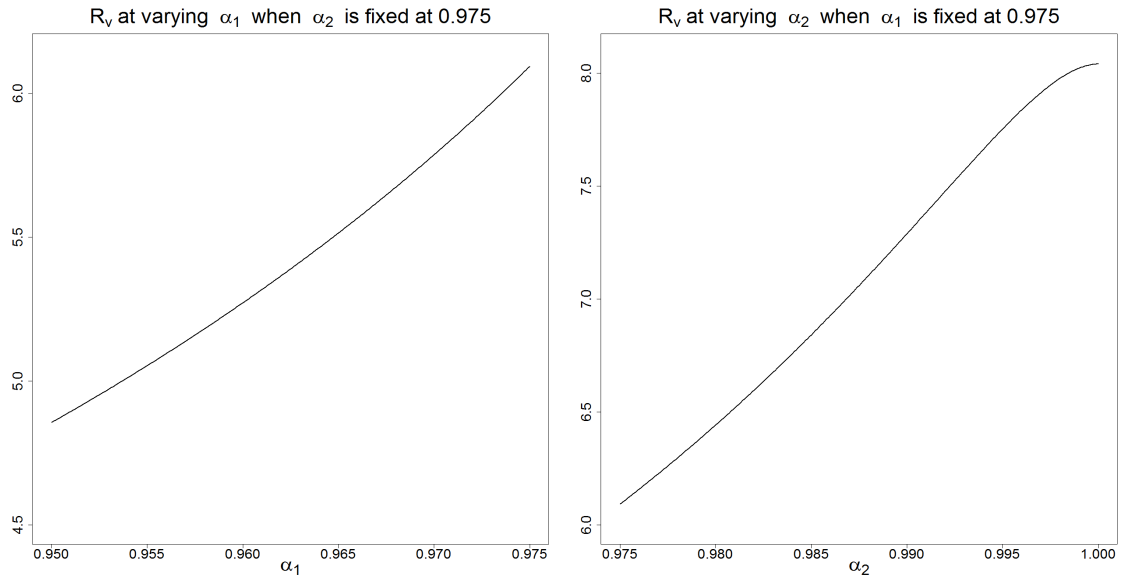

Figure 4: The Berry-Esseen ratio R_v of the uniform spectral test. The left plot shows R_v when we fix $\alpha_2 = 0.975$ and vary α_1 in the interval $(0.95, 0.975)$, and the right plot shows R_v when we fix $\alpha_1 = 0.975$ and vary α_2 in the interval $(0.975, 1)$.

$$\sigma_A^2 = \alpha_1^2(\mu_L - \mu_L^2) + \alpha_2^2(\mu_R - \mu_R^2) + (\mu_{M,2} - \mu_M^2) - 2\mu_M(\alpha_1\mu_L + \alpha_2\mu_R) - 2\alpha_1\alpha_2\mu_L\mu_R\,, \tag{4.13}$$

where

$$\mu_L = F_t(\widehat{F}_t^{-1}(\alpha_1))\,, \tag{4.14}$$

$$\mu_R = 1 - F_t(\widehat{F}_t^{-1}(\alpha_2))\,, \tag{4.15}$$

$$\mu_M = \int_{F_t(\widehat{F}_t^{-1}(\alpha_1))}^{F_t(\widehat{F}_t^{-1}(\alpha_2))} \widehat{F}_t(F_t^{-1}(u))\mathrm{d}u\,, \tag{4.16}$$

$$\mu_{M,2} = \int_{F_t(\widehat{F}_t^{-1}(\alpha_1))}^{F_t(\widehat{F}_t^{-1}(\alpha_2))} \widehat{F}_t(F_t^{-1}(u))^2\mathrm{d}u\,. \tag{4.17}$$

We denote $\overline{W}_v = \frac{1}{n}\sum_{t=1}^n W_{v,t}$, and the approximate finite sample rejection rate of the one-sided uniform spectral test at approximately size κ is given by

$$\begin{aligned}
\mathrm{P}\left(\frac{\sqrt{n}(\overline{W}_v - \mu_v)}{\sigma_v} > \Phi^{-1}(1-\kappa)\right) &= \mathrm{P}\left(\frac{1}{n}\sum_{t=1}^n A_t > \Phi^{-1}(1-\kappa)\frac{\sigma_v}{\sqrt{n}} + \mu_v + \alpha_1\right) \\
&\approx \mathrm{P}\left(Z > \frac{\sqrt{n}}{\sigma_A}\left(\Phi^{-1}(1-\kappa)\frac{\sigma_v}{\sqrt{n}} + \mu_v + \alpha_1 - \mu_A\right)\right) \\
&= 1 - \Phi\left(\frac{\sqrt{n}}{\sigma_A}\left(\Phi^{-1}(1-\kappa)\frac{\sigma_v}{\sqrt{n}} + \mu_v + \alpha_1 - \mu_A\right)\right)\,,
\end{aligned} \tag{4.18}$$

where we have denoted Z to be a standard normal variable.

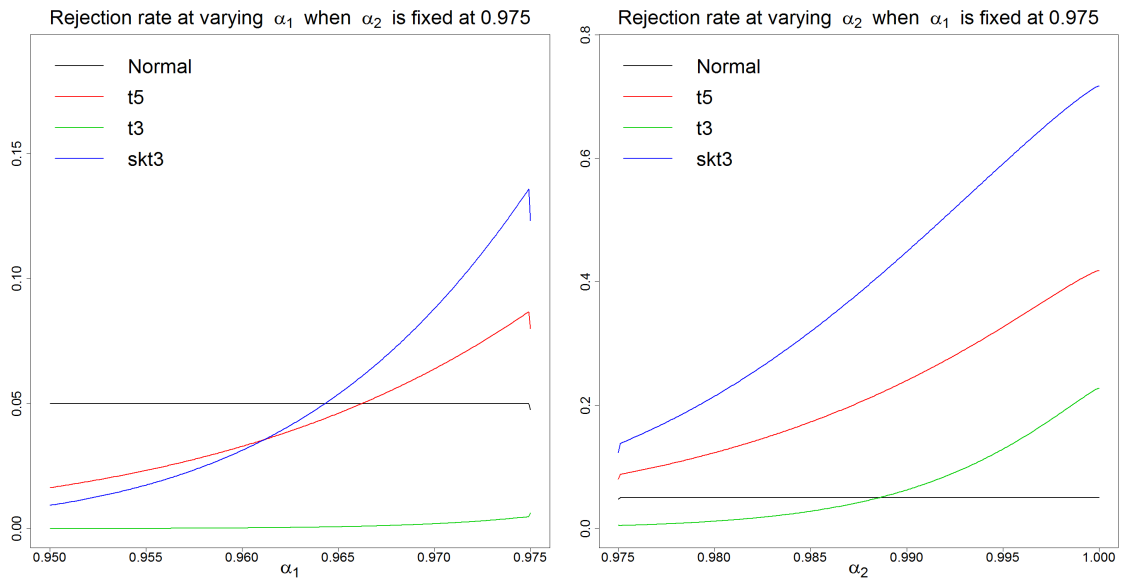

Figure 5: The approximate rejection rate of the one-sided uniform spectral test, when $\widehat{F}_t$ is normal, and F_t is normal, Student-t5, Student-t3, skewed Student-t3 with skewness parameter $\gamma = 1.2$, with $n = 1000$. The left plot shows the rejection rate when we fix $\alpha_2 = 0.975$ and vary α_1 in the interval $(0.95, 0.975)$, and the right plot shows the rejection rate when we fix $\alpha_1 = 0.975$ and vary α_2 in the interval $(0.975, 1)$.

We plot the rejection rate when $\widehat{F}_t$ is normal, and F_t is normal, Student-$t5$, Student-$t3$, skewed Student-$t3$ with skewness parameter $\gamma = 1.2$ in Figure 5, with $n = 1000$. The left plot shows the rejection rate when we fix $\alpha_2 = 0.975$ and vary α_1 in the interval $(0.95, 0.975)$, and the right plot shows the rejection rate when we fix $\alpha_1 = 0.975$ and vary α_2 in the interval $(0.975, 1)$. Notice that for the forecast distributions that we have chosen, when the forecast distribution is misspecified, the rejection rate decreases as we shift α_1 towards 0.5, and increases as we shift α_2 towards 1.

An easier way to understand the rejection rate would be to look at the left plot in Figure 1, which plots $\mathrm{P}(P_t > \alpha) = 1 - \theta_\alpha$, with θ_α as defined in (4.2). Note that the rejection rate seems to depends on the relative difference of the exception rate of the forecast model and the exception rate under the unconditional coverage hypothesis, denoted by $D_r = \int_{\alpha_1}^{\alpha_2} \frac{(1-\theta_u)-(1-u)}{1-u}\,\mathrm{d}u$ (rather than the absolute difference in exception rate, denoted by $D_a = \int_{\alpha_1}^{\alpha_2}(1-\theta_u)-(1-u)\,\mathrm{d}u$, based on findings in Section 4.4.2 later). When $D_r > 0$, the rejection rate will be greater than κ, and conversely, when $D_r < 0$, the rejection rate will be less than κ. Large D_r implies large rejection rate. For example, in the right plot of Figure 5, the rejection rate when F_t is Student-$t3$ and $\widehat{F}_t$ is standard normal (the green line) crosses the black

line at the point $\alpha \neq 0.975$ such that $\int_{0.975}^{\alpha} \frac{u - F_{t_3}(\Phi^{-1}(u))}{1-u} \mathrm{d}u = 0$, where F_{t_3} is the (scaled) Student-$t3$ distribution function.

Summary. The choice of the range determines the trade-off between size and power performance. Nevertheless, it is possible to find a range such that the one-sided uniform spectral test outperforms the one-sided binomial score test both in terms of size and power. For example, in the case when F_t is the (standardized) s$t3$ distribution and $\widehat{F}_t$ is standard normal, the binomial score test at level 99% has Berry-Esseen ratio $R_\alpha = 9.85$ and rejection rate of 75.7% when $n = 1000$. By expanding the range to $\alpha_1 = 0.985$, $\alpha_2 = 0.995$, the one-sided uniform spectral test Berry-Esseen ratio reduces to $R_v = 9.71$ and the rejection rate increases to 81.7% when $n = 1000$.

4.2 In the case when there is parameter estimation error

The style of backtest we implement here is designed to mimic the procedure used in practice where the models are continually updated to use the latest market data. We assumed that the estimated model is updated every 10 steps (which corresponds to two trading week).

4.2.1 Experimental design

In each experiment we generate a dataset of size $n + n_2$ from the true distribution F_t. Similar to Section 4.1.1, we consider the cases when F_t is normal, Student-t distributions with five and three degrees of freedom (denoted by $t5$ and $t3$), and the skewed Student-t distribution with three degrees of freedom and a skewness parameter $\gamma = 1.2$ (denoted s$t3$). We standardized F_t to have zero mean and unit variance. The modeller uses a rolling window of n_2 values to calibrate the forecast distribution $\widehat{F}_t$. We consider 4 possibilities for $\widehat{F}_t$:

The oracle who knows the correct distribution and its exact parameter values.

The good modeller who estimates the correct type of distribution.

The poor modeller who always estimates a normal distribution.

The industry modeller who uses a method known as historical simulation, which we will describe later in Section 4.4.1.

To make the rolling estimation procedure clear, the modeller first use the losses $L_1, \ldots, L_{n_2}$ to calibrate their forecast model $\widehat{F}_{n_2+i}$, and make the realized PIT values $P_{n_2+i} = \widehat{F}_{n_2+i}(L_{n_2+i})$, for $i = 1, \ldots, 10$. The modeller then roll forward 10 steps and use the losses $L_{11}, \ldots, L_{n_2+10}$ to make the realized PIT values $P_{n_2+10+i} = \widehat{F}_{n_2+10+i}(L_{n_2+10+i})$, for $i = 1, \ldots, 10$. Hence, the models are re-estimated $n/10$ times. The experiments are repeated $10,000$ times to determine rejection rates. For cases when computation time are significant, the experiments are repeated $1,000$ times instead. Although the standard error of the rejection rates will be increased by a factor of roughly $\sqrt{10}$, the results should still provide a fairly reliable indication of the performances of the tests. We will divide the analysis to the cases when $\widehat{F}_t$ is parametric and when $\widehat{F}_t$ is non-parametric (the industry modeller).

4.3 When $\widehat{F}_t$ is parametric

The experimental design here is based on Section 4.2.1. Table 9 shows the rejection rate of the two-sided tests similar to those described in Section 4.1.5. We have also included the results for the multinomial tests. Similar to Section 4.1.4, we set $\alpha_1 = \alpha = 0.975$ and further levels are determined by

$$\alpha_j = \alpha + \frac{j-1}{N}(1-\alpha)\,, \quad j = 1, \ldots, N, \;\; N \in \mathbb{N}\,. \tag{4.19}$$

We consider the Pearson and Nass test at $N = 4$, which we denote by P4 and N4, and the discrete probitnormal LRT at $N = 4$ and $N = 8$, which we denote by L4 and L8.

We use the same colouring scheme as previously but some explanations are now required concerning the concepts of size and power. The oracle who knows the correct model should clearly be judged in terms of size. We have decided to judge the good modeller according to the same standards as the oracle. In doing so, we are adopting the same philosophy as Giacomini & White (2006), where we are evaluating the forecasting method, which includes in addition to the forecast model,

the evaluation of the estimation procedure and the data used for estimation. In other words, if the number of data to be used for parameter estimation n_2 is too small, the forecast quality would be poor even if the model is correct, and such a forecasting method should be rejected. The poor modeller should clearly be judged in terms of power. We want to obtain a high rejection rate for this modeller, regardless of whether the parameters of the model are well estimated or not. The results are summarized in Table 9:

Size of the tests. The results for the size of the tests for the oracle and good modeller are summarized in the rows where $\widehat{F}_t$ is "Oracle" and "Good". We observe that for most of the tests, the results for the oracle and the good modeller are in the desired green zone. We note that the sample size in Table 9 is The bispectral tests seem to have a tendency to reject the good modeller with more than 5% probability when $n_2 = 250$.

Power of the tests. The results for the power of the tests for the poor modeller are summarized in the rows where $\widehat{F}_t$ is "Poor". The increased power of the multinomial tests over the binomial test becomes apparent. In particular, L8 performs the best among the multinomial tests. For the spectral and bispectral tests, we observe similar results as in Table 7, where tests that places more emphasis on the tail are more powerful, and the bispectral tests are generally more powerful than the spectral tests. In particular, we almost always obtain good power when using L8, SP.E.200, and the bispectral tests.

n_2	F_t	$\widehat{F}_t$ | test	B99	P4	N4	L4	L8	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	SP.LE.200	PNS	BK
250	Normal	Oracle	4.1	5.8	5.2	5.3	5.3	4.6	4.5	3.9	3.9	4.5	4.9	4.2	5.4
		Good	2.7	4.3	4.1	2.8	2.9	2.7	2.9	4.1	4.3	5.2	5.7	4.8	3.3
		Poor													
	$t5$	Oracle	3.5	5.0	4.9	6.1	5.7	3.9	4.2	4.6	3.8	4.4	4.9	4.7	5.5
		Good	2.7	4.4	4.0	3.4	3.2	2.0	3.0	4.4	4.9	6.2	6.8	6.7	4.6
		Poor	42.6	51.3	50.2	54.4	70.5	39.5	62.4	83.9	80.6	87.0	88.0	91.7	88.2
	$t3$	Oracle	3.4	5.4	4.7	4.4	4.3	4.2	4.2	4.2	4.5	4.4	4.6	4.7	3.8
		Good	3.5	5.7	5.4	3.6	2.9	2.4	3.2	4.2	4.6	5.1	6.1	5.9	3.6
		Poor	49.3	71.8	71.2	81.6	94.1	42.8	72.3	91.0	96.7	99.0	98.7	100.0	99.6
	$st3$	Oracle	3.5	5.2	5.2	5.6	4.8	5.0	4.6	3.8	4.2	4.7	4.6	4.0	4.9
		Good	3.3	4.2	4.2	3.5	3.2	2.8	3.6	5.2	4.1	5.8	6.1	6.5	3.6
		Poor	92.1	96.5	96.2	97.1	99.4	91.3	97.6	99.7	99.6	99.9	99.8	99.9	99.9
500	Normal	Oracle	2.9	4.8	4.4	4.8	4.5	4.5	4.8	4.3	4.5	4.5	4.8	4.2	4.5
		Good	3.0	4.4	4.2	2.9	2.1	2.8	3.1	3.2	2.8	3.3	4.1	3.9	3.0
		Poor													
	$t5$	Oracle	3.6	5.2	5.1	5.9	5.7	5.3	4.7	5.0	5.1	4.6	4.2	5.4	5.1
		Good	2.2	4.4	4.2	2.4	3.4	2.0	3.1	3.6	3.8	3.7	4.1	3.5	2.9
		Poor	36.5	43.6	42.7	48.7	68.5	35.2	57.2	77.9	74.3	83.4	84.5	89.4	84.1
	$t3$	Oracle	2.8	4.6	4.4	5.0	5.0	4.9	3.9	4.1	5.2	4.4	4.4	4.6	4.1
		Good	1.9	4.0	3.9	3.9	3.3	1.7	2.0	2.5	2.9	3.1	3.2	3.4	2.6
		Poor	41.9	70.5	69.6	82.9	92.8	35.3	62.4	84.8	94.9	96.7	96.6	98.0	97.9
	$st3$	Oracle	3.4	5.1	4.8	4.6	4.3	4.8	4.1	4.5	4.3	5.6	5.9	5.3	4.5
		Good	2.2	4.0	3.7	2.9	2.7	1.9	2.0	2.8	3.3	3.8	4.0	3.8	2.7
		Poor	83.7	93.7	93.3	96.0	99.3	80.9	94.3	99.0	99.8	100.0	99.9	100.0	100.0

Table 9: Rejection rates for the various VaR estimation methods using various two-sided tests. Models are refitted after 10 simulated values and backtest length is 1000. Results are based on 1000 replications, with $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$.

4.4 When $\widehat{F}_t$ is non-parametric

4.4.1 Historical simulation method

We denote the historically simulated losses at time t by $\mathcal{S}_t = \{L_{t,1}, \ldots, L_{t,n_2}\}$, where n_2 refers to amount of data to be used for model calibration. Some banks construct the HS model using the standard empirical distribution function

$$\widehat{F}_t(x) = \frac{1}{n_2} \sum_{j=1}^{n_2} I_{\{L_{t,j} \leqslant x\}} \,. \tag{4.20}$$

It is also common to use a linear interpolation method, which we will refer to as the HS-Linear method. Let $L_{t,(1)} < \cdots < L_{t,(n_2)}$ denote the order statistics of $\mathcal{S}_t$. For $j = 1, \ldots, n_2 - 1$, the empirical distribution function is given by

$$\widehat{F}_t(x) = \frac{j}{n_2} \frac{L_{t,(j+1)} - x}{L_{t,(j+1)} - L_{t,(j)}} + \frac{j+1}{n_2} \frac{x - L_{t,(j)}}{L_{t,(j+1)} - L_{t,(j)}}, \quad L_{t,(j)} \leqslant x \leqslant L_{t,(j+1)} \,. \tag{4.21}$$

In both cases we have $\widehat{F}_t(x) = 0$ for $x < L_{t,(1)}$ and $\widehat{F}_t(x) = 1$ for $x > L_{t,(n_2)}$ so that it is not possible to assign meaningful probabilities outside the range of the data.

Another disadvantage of the historical simulation method is that, even within the range of the data, it does not give good estimates of the tail of F_t unless the the window size n_2 is very large. To understand this, we construct the empirical estimator of the ES at level α at time t using the PIT-based VaR exception, given by

$$\widehat{\mathrm{ES}}_{\alpha,t} = \frac{\sum_{j=1}^{n_2} L_{t,j} I_{\{P_{t,j} > \alpha\}}}{\sum_{j=1}^{n_2} I_{\{P_{t,j} > \alpha\}}}, \quad P_{t,j} = \widehat{F}_t(L_{t,j}) \,. \tag{4.22}$$

To understand this estimator, recall the relationship (3.28) between VaR exceptions and realized PIT values. Note that both the HS and HS-Linear method in (4.20) and (4.21) will give the same empirical expected shortfall estimator in (4.22).

Table 10 shows the bias and mean absolute error (MAE) of the empirical estimator $\widehat{\mathrm{ES}}_{0.975,t}$ for different values of n_2 and different underlying distribution F_t. We observe that there is always a negative bias, which decreases with n_2 and increases with the heaviness of the tail of F_t. The MAE also decreases with n_2 and increases with the tail of F_t. The final three columns show the estimated probability (expressed as a percentage) of underestimating $\mathrm{ES}_{0.975}$ by 10%, 25% or 33%. These probabilities are

considerable for $n_2 = 250$. The results suggest that companies should be discouraged from using short windows for historical simulation calibration.

n_2	F_t | results	Bias	MAE	By10	By25	By33
250	Normal	-3.2	7.1	20.6	0.2	0.0
	$t5$	-4.5	12.8	40.0	5.9	0.7
	$t3$	-5.4	19.5	49.7	19.1	6.7
	s$t3$	-6.5	20.9	53.2	22.5	9.2
500	Normal	-1.4	4.9	7.0	0.0	0.0
	$t5$	-1.9	9.2	24.6	0.6	0.0
	$t3$	-2.1	14.4	37.9	6.2	0.8
	s$t3$	-2.7	15.4	40.3	8.3	1.4
1000	Normal	-0.4	3.5	1.3	0.0	0.0
	$t5$	-0.6	6.6	12.0	0.0	0.0
	$t3$	-0.4	10.5	24.1	0.9	0.1
	s$t3$	-0.6	11.4	27.9	1.4	0.0

Table 10: Bias and mean absolute error (MAE) (both expressed as percentages) of standard empirical estimator of 97.5% expected shortfall for different sample sizes and different distributions. By10, By25 and by33 give percentage of estimates underestimating expected shortfall by 10%, 25% or 33.3% respectively. Results are based on 10000 replications.

4.4.2 Rejection rate for historical simulation method

The experimental design here is based on Section 4.2.1. Table 11 and Table 12 shows the rejection rate of the two-sided multinomial, spectral and bispectral tests similar to those described in Section 4.3, when the forecast distribution is the empirical distribution in (4.20) and (4.21).

We have added two additional tests:

Two-sided zero mean test for expected shortfall which we denote by M.true, where we compare the empirical estimator of expected shortfall in (4.22) with the "true" expected shortfall, denoted by $\mathrm{ES}_{\alpha_1,t}$. The Z-test statistic is

$$Z = \frac{\sum_{t=1}^{n} d_t}{\sqrt{\sum_{t=1}^{n} d_t^2}}, \quad d_t = \mathrm{ES}_{\alpha_1,t} - \widehat{\mathrm{ES}}_{\alpha_1,t}. \tag{4.23}$$

The two-sided hypothesis is

$$H_0 : \ \mathrm{E}(d_t) = 0 \ \text{ vs. } \ H_1 : \mathrm{E}(d_t) \neq 0, \tag{4.24}$$

and we reject H_0 when $Z^2 > F^{-1}_{\chi^2_1}(1-\kappa)$ to obtain a test of approximately size κ.

Clearly, this test requires the use of the "true" expected shortfall, which is not known in practice. This test is only valid from an internal model validation point of view, where the forecaster has a system to construct forecast models given arbitrary data sets. The forecaster then feeds simulated data to this system to estimate $\widehat{\mathrm{ES}}_{\alpha_1,t}$, when $\mathrm{ES}_{\alpha_1,t}$ is known, and try to undertand the ability of the system to produce accurate expected shortfall estimates. See Jarvis et al. (2016) for an example of this type of model validation technique.

Two-sided test for expected shortfall as described in Section 3.3, which we denote by M.est, where we have used the empirical VaR estimate

$$\widehat{\mathrm{VaR}}_{\alpha,t} = \inf\{L_{t,j} : \widehat{F}_t(L_{t,j}) > \alpha\}\,, \quad j = 1, \ldots, n_2\,. \tag{4.25}$$

The HS method is acceptable provided that enough data is used. However it is less easy to say what is enough data because that depends on how heavy the tails of the underlying distribution is. In view of the results in Table 10, and to keep things simple we have made the arbitrary decision to apply power colouring.

We will refer to the model in (4.20) as the HS model, and (4.21) as the HS-Linear model. The results are summarized in Table 11 and Table 12. We observed that the binomial test B99 have no power in rejecting the HS model, and a little power in rejecting the HS-Linear model when $n_2 = 250$. The same holds true for the discrete probitnormal LRT, where L8 is more powerful than B99 in rejecting the HS-Linear model when $n_2 = 250$. Both Pearson (P4) and Nass (N4) test have some power in rejecting the HS and HS-Linear model when $n_2 = 250$, but not when $n_2 = 500$. As for the spectral tests, SP.E.200 which places most emphasis in the tail performs the best, and have decent power in rejecting the HS-Linear model when $n_2 = 250$, and some power in rejecting the HS model when $n_2 = 250$ and the HS-Linear model when $n_2 = 500$. For the bispectral tests, PNS performs the best, as it is able to reject HS and HS-Linear model when $n_2 = 250$ with high power, especially when n is large ($n = 1000, 2000$). It have some power in rejecting rejecting the HS-Linear model when $n_2 = 500$, but is unable to reject the HS model when $n_2 = 500$.

The M.true test in (4.23) has rather good power in rejecting both HS ands HS-Linear model for both $n_2 = 250$ and $n_2 = 500$, which is as expected given the results in Table 10. In contrast, M.est in (3.27) has very poor power in rejecting both HS ands HS-Linear model for both $n_2 = 250$ and $n_2 = 500$. This is because we are attempting to compare the mean of the empirical expected shortfall with the mean of the "realized" empirical expected shortfall. Given that the models are updated every 10 steps, these values are unlikely to differ by much in the static backtest setting.

In summary, excluding M.true (as it is not possible to use this test in a regulatory setting), the PNS test performs the best in detecting poorly calibrated HS and HS-Linear models. However, all tests, including PNS, have difficulty in rejecting the HS model when $n_2 = 500$.

From Table 11 and Table 12, we notice that the HS-Linear model in (4.21) is rejected more strongly than the HS model in (4.20). Also, the rejection rate does not depend on the data generating distribution F_t. This is because regardless of F_t, the exception rate of the HS and HS-Linear model at level α, denoted by $1-\theta_\alpha = \mathrm{P}\left(P_t > \alpha\right)$, with θ_α as defined in (4.2), remains the same, and is shown in Figure 6. For reference, we have also included the exception rate when $\widehat{F}_t$ is standard normal, and F_t is (scaled) Student-$t5$, which we will simply label as $t5$.

From the graph, it is obvious that the HS-Linear model will underestimates the tail more often than the HS model. We also observe that for the HS-Linear model, even though $(1-\theta_u)-(1-u)$ remains constant in the region $u \in (\alpha_1, \alpha_2)$, spectral test with increasing emphasis on the tail is more powerful (SP.E.200 is more powerful than SP.U), which means that the rejection rate of spectral test is likely to depend on the relative difference in exception rate $\int_{\alpha_1}^{\alpha_2} \frac{(1-\theta_u)-(1-u)}{1-u}\,\mathrm{d}u$ rather than the absolute difference in exception rate $(\int_{\alpha_1}^{\alpha_2}(1-\theta_u)-(1-u)\,\mathrm{d}u)$ between the forecast model and the true model.

n_2	F_t	n | test	B99	P4	N4	L4	L8	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	SP.LE.200	PNS	BK	M.true	M.est
250	Normal	250	5.6	7.4	6.6	4.3	3.9	5.1	6.9	11.4	7.3	10.2	12.1	21.3	12.5	96.6	3.7
		500	3.7	6.1	5.8	2.0	1.5	2.7	5.1	9.9	5.0	9.7	13.0	24.6	14.4	93.7	0.3
		1000	2.7	10.2	9.7	1.2	1.0	1.8	5.3	16.6	8.2	16.1	21.6	41.2	29.6	93.6	0.6
		2000	5.6	25.4	24.2	1.4	0.8	2.7	9.3	33.0	14.5	32.1	42.6	72.8	62.9	94.6	4.9
	$t5$	250	6.2	8.0	7.5	3.9	4.4	5.4	7.1	11.2	8.0	11.5	13.5	22.2	14.3	95.9	7.6
		500	2.8	6.2	5.5	1.9	1.9	2.3	4.4	9.0	5.9	10.1	14.3	25.4	15.8	92.7	1.0
		1000	2.4	11.4	10.9	1.4	1.3	2.0	4.7	15.6	7.2	16.3	20.8	39.5	29.6	93.9	0.1
		2000	4.5	25.8	25.0	1.6	0.6	2.4	9.8	34.2	13.5	32.4	43.5	72.1	62.8	93.3	0.3
	$t3$	250	5.7	6.9	6.3	4.0	4.5	4.9	7.3	12.1	8.0	12.2	14.3	22.7	15.0	95.5	8.2
		500	2.4	5.8	5.1	1.1	1.2	2.5	4.9	12.2	6.1	12.6	16.0	27.8	17.5	94.0	1.8
		1000	2.6	10.6	10.0	1.6	0.7	2.7	5.9	16.7	6.5	16.5	22.8	41.8	30.8	93.6	0.2
		2000	4.6	22.7	21.7	1.0	0.3	2.2	8.5	32.3	12.9	33.2	42.5	72.0	62.7	93.2	0.3
	$st3$	250	6.1	8.6	8.2	4.3	4.5	6.8	8.3	13.6	8.9	13.9	16.5	26.4	17.0	95.0	8.1
		500	2.3	6.3	5.4	1.6	1.2	3.1	5.3	11.6	5.6	11.1	15.7	28.0	17.4	93.0	1.7
		1000	3.5	11.8	11.1	0.8	0.7	2.7	5.5	17.2	7.1	18.1	22.7	43.4	31.7	93.7	0.1
		2000	4.7	26.0	25.4	1.0	0.5	2.6	11.0	36.4	13.9	33.5	43.1	74.0	64.8	92.6	0.0
500	Normal	250	3.7	5.2	5.0	6.3	6.2	6.1	6.6	7.7	5.9	6.9	8.0	11.9	9.1	96.1	9.1
		500	1.6	3.7	3.0	2.9	2.8	3.8	4.8	6.3	3.7	5.0	6.3	10.7	6.9	95.6	3.2
		1000	0.2	2.3	2.1	1.6	0.2	1.0	1.5	3.6	1.1	2.8	3.4	9.2	5.1	93.5	0.2
		2000	0.1	2.4	2.2	1.6	0.0	0.3	0.5	3.1	1.0	3.4	4.9	13.0	8.7	93.4	0.0
	$t5$	250	3.3	5.3	4.8	4.8	5.2	4.6	6.0	8.6	6.0	8.2	8.6	13.0	8.0	96.8	13.2
		500	1.8	4.4	4.1	3.6	3.2	4.2	4.7	6.5	4.4	6.3	7.1	11.0	7.6	96.0	5.5
		1000	0.2	2.6	2.4	2.0	0.8	0.7	1.1	4.2	1.4	3.6	4.4	10.3	5.9	93.9	0.5
		2000	0.1	2.2	2.1	1.7	0.3	0.1	0.5	2.5	0.9	2.5	4.5	13.5	8.2	94.9	0.0
	$t3$	250	2.5	5.8	5.7	5.7	5.6	5.1	5.7	8.2	7.0	9.6	10.4	12.8	10.1	96.8	16.2
		500	1.7	2.9	2.6	3.5	2.0	2.8	3.6	6.0	3.9	5.4	6.8	10.1	6.5	96.5	6.4
		1000	0.3	1.9	1.7	1.8	0.8	0.8	1.5	3.9	1.4	2.7	3.5	8.3	5.0	95.6	0.8
		2000	0.2	2.2	2.1	1.5	0.3	0.0	0.4	2.4	0.8	2.2	4.0	12.3	8.6	93.3	0.1
	$st3$	250	2.8	6.0	5.5	6.2	6.2	4.8	4.9	8.2	5.9	8.5	9.8	12.5	10.1	96.0	17.0
		500	2.2	4.5	3.9	3.9	2.1	3.7	4.8	6.3	4.0	5.4	6.6	11.0	7.0	95.3	6.8
		1000	0.5	2.1	2.1	1.7	1.2	1.1	2.1	3.7	2.0	3.0	3.6	9.1	5.1	94.9	1.4
		2000	0.0	2.4	2.3	1.5	0.3	0.3	0.3	3.0	1.0	2.6	4.5	12.8	8.8	94.9	0.2

Table 11: Rejection rates of the various two-sided tests for the HS method with empirical distribution in (4.20). Models are refitted after 10 simulated values. Results are based on 1000 replications, with $\alpha_1 = 0.975$ *and* $\alpha_2 = 0.9995$.

n_2	F_t	n | test	B99	P4	N4	L4	L8	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	SP.LE.200	PNS	BK	M.true	M.est
250	Normal	250	10.7	8.0	7.2	5.7	8.3	9.9	15.4	22.1	14.7	20.6	22.3	31.6	18.9	96.6	3.7
		500	7.9	6.8	5.9	4.5	7.0	9.1	15.0	28.4	15.8	25.8	30.8	44.6	29.2	93.7	0.3
		1000	10.9	8.9	8.3	6.9	15.4	13.5	27.4	55.5	28.8	49.1	58.3	74.9	62.0	93.6	0.6
		2000	34.4	21.1	20.5	20.2	44.5	37.2	69.6	94.2	64.8	88.7	91.7	98.5	96.1	94.6	4.9
	$t5$	250	10.0	6.8	6.2	5.4	7.5	10.1	13.3	20.2	14.6	19.1	21.1	29.6	18.1	95.9	7.6
		500	6.8	6.8	5.9	3.9	7.0	8.1	14.0	27.0	14.6	25.7	31.0	42.8	29.3	92.7	1.0
		1000	9.9	9.8	9.2	7.2	13.2	12.9	26.9	52.6	27.0	45.4	52.2	73.6	58.5	93.9	0.1
		2000	29.1	18.7	17.9	17.3	36.0	34.6	64.6	91.2	61.0	85.6	89.5	98.1	95.2	93.3	0.3
	$t3$	250	9.4	7.0	6.4	5.2	7.9	9.9	14.0	21.6	14.2	20.7	23.3	31.7	19.0	95.5	8.2
		500	6.7	6.6	5.4	3.1	6.3	7.8	13.7	28.8	14.8	27.2	30.8	42.3	30.3	94.0	1.8
		1000	10.8	8.1	7.8	6.2	11.6	13.6	26.3	52.4	25.2	45.8	52.7	72.4	58.2	93.6	0.2
		2000	28.6	16.5	15.7	15.7	33.4	35.1	65.4	90.9	58.9	82.4	86.7	96.9	94.9	93.2	0.3
	$st3$	250	11.0	8.3	7.9	6.4	8.0	10.8	15.1	21.8	15.2	21.8	24.7	34.7	20.8	95.0	8.1
		500	7.1	6.0	4.7	3.7	6.0	8.7	14.4	28.2	14.8	25.4	30.6	43.5	31.1	93.0	1.7
		1000	10.2	9.2	8.6	7.0	11.6	13.8	26.3	52.6	27.0	44.2	51.3	71.2	58.1	93.7	0.1
		2000	31.3	20.1	19.0	17.8	37.2	36.9	66.4	91.2	60.7	83.8	88.4	98.0	95.4	92.6	0.0
500	Normal	250	7.8	6.0	5.6	6.6	6.6	7.9	9.2	11.7	10.3	11.5	12.7	17.2	9.8	96.1	9.1
		500	5.1	4.0	3.5	3.3	5.0	6.6	8.7	11.7	7.6	11.3	12.8	18.2	11.2	95.6	3.2
		1000	3.4	2.6	2.6	1.4	3.2	3.3	6.3	12.9	5.9	10.9	13.8	22.8	13.0	93.5	0.2
		2000	4.5	3.0	2.8	1.4	4.1	2.7	6.8	25.1	8.1	20.4	27.5	41.8	31.2	93.4	0.0
	$t5$	250	6.3	6.5	6.0	5.2	6.8	6.6	9.8	13.9	9.4	13.4	14.7	18.0	11.9	96.8	13.2
		500	5.1	5.7	5.0	3.9	4.2	7.1	8.9	13.0	7.1	10.9	12.2	17.1	11.7	96.0	5.5
		1000	2.7	2.6	2.5	1.5	2.6	3.0	6.1	14.0	5.4	9.8	13.3	22.1	13.5	93.9	0.5
		2000	4.5	3.6	3.2	1.0	2.6	2.4	7.0	22.3	7.1	18.4	23.7	38.5	27.0	94.9	0.0
	$t3$	250	6.7	6.3	6.2	5.4	7.0	7.5	8.7	13.1	10.0	12.9	15.0	17.9	12.0	96.8	16.2
		500	4.7	3.7	3.3	3.8	3.7	6.0	8.3	11.5	6.6	10.4	12.3	16.1	10.7	96.5	6.4
		1000	3.7	2.7	2.7	1.8	2.1	3.2	6.7	13.6	5.4	10.0	12.8	19.0	11.6	95.6	0.8
		2000	5.4	3.3	3.0	1.4	2.7	3.1	8.3	21.4	5.8	17.3	22.0	36.3	25.8	93.3	0.1
	$st3$	250	6.2	5.7	5.1	6.6	6.3	6.4	7.8	12.0	8.8	12.4	13.9	16.2	11.7	96.0	17.0
		500	5.9	5.0	4.5	3.8	3.8	6.7	7.3	11.8	8.4	10.9	11.8	17.4	10.3	95.3	6.8
		1000	3.8	3.2	2.9	1.8	2.6	3.7	6.4	12.2	5.5	9.4	11.6	20.4	11.2	94.9	1.4
		2000	5.2	2.8	2.8	1.9	2.9	2.2	7.0	21.6	7.0	16.0	21.1	38.4	26.6	94.9	0.2

Table 12: Rejection rates of the various two-sided tests for the HS-Linear method with empirical distribution in (4.21). Models are refitted after 10 simulated values. Results are based on 1000 replications, with $\alpha_1 = 0.975$ *and* $\alpha_2 = 0.9995$.

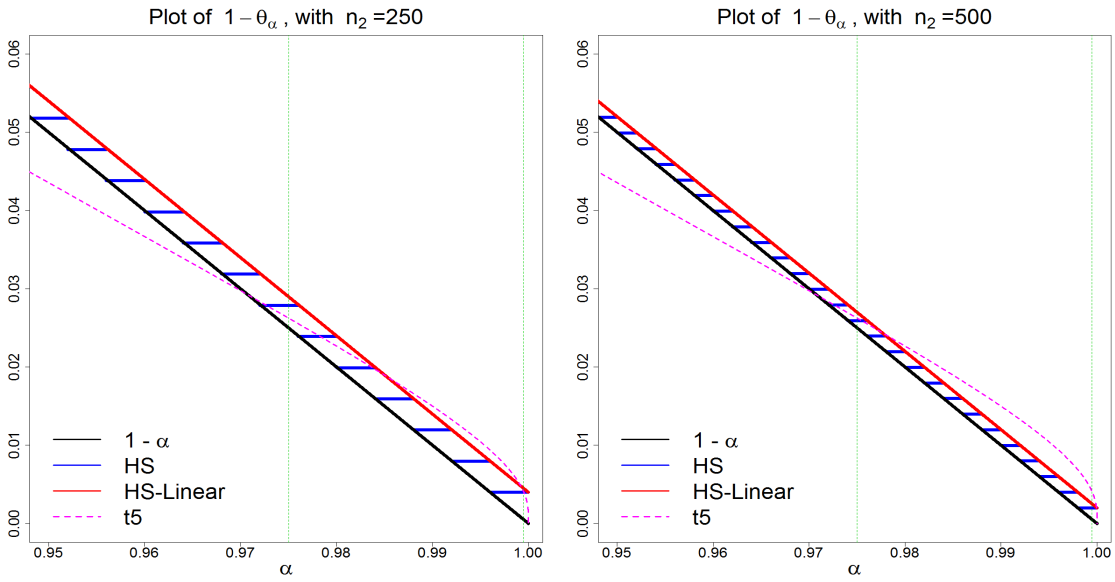

Figure 6: The exception rate $1-\theta_\alpha$ *of the HS and HS-Linear model, when* $n_2 = 250$ *and* $n_2 = 500$*. For reference, we have also included the exception rate when* $\widehat{F}_t$ *is standard normal, and* F_t *is (scaled) Student-t5, which we will simply label as t5. The black line represents the case when* $\theta_\alpha = \alpha$*, and the green vertical lines represents the levels* $\alpha = 0.975$ *and* $\alpha = 0.9995$*.*

Chapter 5 Explicit testing for serial independence of $W_{v,t}$

Recall from Section 3.4 that for a series of losses (L_t) with conditional distribution given information up to time $t-1$

$$F_t(x) = \mathrm{P}\ (L_t \leqslant x \mid \mathcal{F}_{t-1})\ , \tag{5.1}$$

the series of realized PIT values, denoted by (P_t), with $P_t = \widehat{F}_t(L_t)$ should behave as iid standard uniform variables when the forecast distribution $\widehat{F}_t = F_t$.

When the forecast distribution inadequately models the dynamics of the losses (L_t), (P_t) and hence $(W_{v,t})$ can have structural shifts at unknown dates. Previously, we have tested for serial independence implicitly using the Z-test, where in the construction of the Z-test statistic, we have assumed that $\mathrm{var}(\sum_{t=1}^n W_{v,t}) = \sum_{t=1}^n \mathrm{var}(W_{v,t})$. When the forecast model fails to capture, say the volatility clustering effect of the losses, $\mathrm{cov}(W_{v,t}, W_{v,t-k})$ is likely to be positive for small k, leading to $\sum_{t=1}^n \mathrm{var}(W_{v,t}) \leqslant \mathrm{var}(\sum_{t=1}^n W_{v,t})$, hence the test statistic will be larger and we observe a larger rejection rate.

In this chapter, we introduce several methods to test for serial independence of $W_{v,t}$ explicitly, as well as summarize some of the more popular existing methods to test for serial independence.

5.1 Portmanteau tests using autocorrelation function

The simplest way to test for serial independence of $(W_{v,t})$ is to test the autocorrelation function (acf) of $(W_{v,t})$. We will follow closely the work in Brockwell & Davis (1991) and Brockwell & Davis (2003).

Let (X_t) be a time series with $\mathrm{E}\,(X_t^2) < \infty$. (X_t) is weakly stationary if the mean $\mathrm{E}\,(X_t) = \mu$ is independent of t and the autocovariance function $\gamma(h) = \mathrm{cov}(X_{t+h}, X_t)$ is independent of t and h. We denote by $\rho(h) = \frac{\gamma(h)}{\gamma(0)}$ the acf of (X_t).

Let $(x_1, \ldots, x_n)$ be a realization of the time series (X_t). We denote by $\bar{x} = \frac{1}{n}\sum_{t=1}^n x_t$

the sample mean. The sample autocovariance function is

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \bar{x})(x_t - \bar{x})\,, \tag{5.2}$$

and the sample autocorrelation function is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -n < h < n\,. \tag{5.3}$$

By the Wold's decomposition, we can re-write the weakly stationary time series (X_t) as

$$X_t - \mu = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}\,, \tag{5.4}$$

where (Z_t) is a series of iid random variable with mean zero and variance σ^2, and $\sum_{j=-\infty}^{\infty}|\psi_j| < \infty$. We denoted the acf and sample acf vector by $\boldsymbol{\rho}(h) = \big(\rho(1),\ldots,\rho(h)\big)$ and $\hat{\boldsymbol{\rho}}(h) = \big(\hat{\rho}(1),\ldots,\hat{\rho}(h)\big)$. Under the condition that $\sum_{j=-\infty}^{\infty}\psi_j^2|j| < \infty$, $\hat{\boldsymbol{\rho}}(h)$ is asymptotically $N_h\big(\boldsymbol{\rho}(h), n^{-1}\Sigma\big)$ distributed, where Σ is the covariance matrix whose (i,j) element is given by Barlett's formula, with

$$\Sigma_{ij} = \sum_{k=1}^{\infty}\big(\rho(k+i)+\rho(k-i)-2\rho(i)\rho(k)\big)\big(\rho(k+j)+\rho(k-j)-2\rho(j)\rho(k)\big)\,. \tag{5.5}$$

See Brockwell & Davis (1991) Theorem 7.2.2. Under the null hypothesis of serial independence, $\rho(l) = 0$ for $|l| > 0$, hence Σ becomes the identity matrix and $\hat{\rho}(1),\ldots,\hat{\rho}(h)$ are asymptotically iid $N(0, n^{-1})$ distributed.

Using the above results, Box & Pierce (1970) considered the test statistic

$$S_{\text{B-P}} = n\sum_{k=1}^{h}\hat{\rho}(k)^2\,, \tag{5.6}$$

which is asymptotically χ^2_h distributed. Ljung & Box (1978) did a refinement to the Box-Pierce test, with test statistic given by

$$S_{\text{L-B}} = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}(k)^2}{n-k}\,, \tag{5.7}$$

which is also asymptotically χ^2_h distributed.

5.2 Tests based on martingale difference property

In this section we consider testing for the serial independence of the weighted realized PIT values within a regression or conditional framework. Let $(W_{v,t})$ denote the

sequence of weighted realized PIT and let $(\mathcal{F}_t)$ denote the filtration generated by the realized PIT values, i.e. $\mathcal{F}_t = \sigma(P_1, \ldots, P_t)$. Let $(\tilde{W}_{v,t})$ denote the sequence of weighted realized PIT values centered at zero under the null hypothesis (3.30), with $\tilde{W}_{v,t} = W_{v,t} - \mu_v$.

We test for the martingale difference (MD) property

$$H_0 : \quad \mathrm{E}\,(\tilde{W}_{v,t} \mid \mathcal{F}_{t-1}) = 0 \tag{5.8}$$

which is necessary for $(W_{v,t})$ to be an iid sequence with mean μ_v.

5.2.1 Conditional spectral test

If $(\tilde{W}_{v,t})$ is an MD sequence then, for any $\mathcal{F}_{t-1}$-measurable random variable r_{t-1} we must have $\mathrm{E}\,(r_{t-1}\tilde{W}_{v,t} \mid \mathcal{F}_{t-1}) = 0$. We form the $(h+1)$-dimensional lagged vector

$$\boldsymbol{r}_{t-1} = \big(1, f(P_{t-1}), \ldots, f(P_{t-h})\big)^T \tag{5.9}$$

for some function f and base our test on the vector-valued process $\boldsymbol{V}_t = \boldsymbol{r}_{t-1}\tilde{W}_{v,t}$ for $t = h+1, \ldots, n$. Under the null hypothesis (5.8) the process $(\boldsymbol{V}_t)$ is a MD sequence satisfying

$$\mathrm{E}\,(\boldsymbol{V}_t \mid \mathcal{F}_{t-1}) = \mathbf{0}\,, \quad t = h+1, \ldots, n\,. \tag{5.10}$$

Let $\overline{\boldsymbol{V}} = (n-h)^{-1} \sum_{t=h+1}^{n} \boldsymbol{V}_t$ and let $\hat{\Sigma}_V$ denote a consistent estimator of $\Sigma_V = \mathrm{cov}(\boldsymbol{V}_t)$.

which was developed for comparing forecasting methods can be applied in this context. Giacomini & White (2006) show that under very weak assumptions, for large enough n and fixed h,

$$S = (n-h)\, \overline{\boldsymbol{V}}^T\, \hat{\Sigma}_V^{-1}\, \overline{\boldsymbol{V}} \sim \chi^2_{h+1}\,. \tag{5.11}$$

Giacomini & White (2006) have used the estimator $\hat{\Sigma}_V = (n-h)^{-1} \sum_{t=h+1}^{n} \boldsymbol{V}_t\, \boldsymbol{V}_t^T$.

We propose a different estimator for Σ_V. Under the null hypothesis (5.8) and the additional assumption that $\mathrm{E}\,(\tilde{W}^2_{v,t} \mid \mathcal{F}_{t-1}) = \sigma_v^2$ for all t, we can compute that

$$\begin{aligned}
\Sigma_V = \mathrm{E}\,\big(\mathrm{cov}(\boldsymbol{V}_t \mid \mathcal{F}_{t-1})\big) = \mathrm{E}\,\big(\mathrm{E}\,(\boldsymbol{V}_t\, \boldsymbol{V}_t^T \mid \mathcal{F}_{t-1})\big) &= \mathrm{E}\,\Big(\boldsymbol{r}_{t-1}\, \boldsymbol{r}_{t-1}^T \mathrm{E}\,(\tilde{W}^2_{v,t} \mid \mathcal{F}_{t-1})\Big) \\
&= \sigma_v^2 \mathrm{E}\,\big(\boldsymbol{r}_{t-1}\, \boldsymbol{r}_{t-1}^T\big) \qquad (5.12) \\
&= \sigma_v^2\, \mathrm{diag}(1, \sigma_r^2, \ldots, \sigma_r^2)\,, \quad (5.13)
\end{aligned}$$

where $\sigma_r^2 = \mathrm{E}\left(f(P_t)^2\right)$.

There are a number of possibilities for the choice of f. The simplest choice would be to set $f(P_t) = \tilde{W}_{v,t}$ so that $\sigma_r^2 = \sigma_v^2$. In the case where $\mathrm{E}\left(f(P_t)^2\right)$ is difficult to compute, we consider the hybrid approach in which we use the estimator

$$\hat{\Sigma}_V = \frac{\sigma_v^2}{n-h} \sum_{t=h+1}^{n} \boldsymbol{r}_{t-1}\, \boldsymbol{r}_{t-1}^T \tag{5.14}$$

based on (5.12) and the value of σ_v^2 under the null hypothesis (3.30). The latter approach gives a test statistic that is a generalization of the out-of-sample dynamic quantile test statistic proposed by Engle & Manganelli (2004).

To see this let M be the $(n-h) \times (h+1)$ matrix whose rows are given by $\boldsymbol{r}_{t-1}$ for $t = h+1, \ldots, n$. Let $\tilde{\boldsymbol{W}}_v = (\tilde{W}_{v,h+1}, \ldots, \tilde{W}_{v,n})^T$. It is easy to check that $\hat{\Sigma}_V = \frac{\sigma_v^2}{n-h} \sum_{t=h+1}^{n} \boldsymbol{r}_{t-1}\, \boldsymbol{r}_{t-1}^T = \frac{\sigma_v^2}{n-h} M^T M$ and $\overline{\boldsymbol{V}} = \frac{1}{n-h} M^T \tilde{\boldsymbol{W}}_v$ so that (5.11) may be rewritten as

$$\sigma_v^{-2} \tilde{\boldsymbol{W}}_v^T M (M^T M)^{-1} M^T \tilde{\boldsymbol{W}}_v \sim \chi_{h+1}^2 \,. \tag{5.15}$$

We will refer to the case where the measure corresponds to point mass at some level α as the conditional binomial test. In this case, the weight measure $v = \delta_\alpha$, and $\tilde{W}_{v,t} = I_{\{P_t > \alpha\}} - \mu_v$, with $\mu_v = 1-\alpha$ and $\sigma_v^2 = \alpha(1-\alpha)$. If we use the lagged variables $f(P_t) = \tilde{W}_{v,t}$, (5.15) is the statistic proposed by Engle & Manganelli (2004).

Another way of deriving (5.15) is to consider the regression model

$$\tilde{W}_{v,t} = \beta_0 + \sum_{i=1}^{h} \beta_i f(P_{t-i}) + \epsilon_t, \quad t = h+1, \ldots, n \,, \tag{5.16}$$

and assume that $\mathrm{var}(\epsilon_t \mid \mathcal{F}_{t-1}) = \sigma_v^2$ for all t so that the errors are homoscedastic with (known) variance σ_v^2. The matrix M is the design matrix in (5.16) and the expression (5.15) describes the test for the null hypothesis

$$H_0 : \quad \beta_0 = \beta_1 = \cdots = \beta_h = 0 \tag{5.17}$$

based on the asymptotic normality of the least squares estimator of the unknown parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_h)^T$.

5.2.2 Conditional bispectral test

Suppose we have two sets of weighted realized PIT values $(W_{v_1,t}, W_{v_2,t})$ for $t = 1, \ldots, n$. We now form the vector $\boldsymbol{V}_t$ of length $h+2$ given by

$$\boldsymbol{V}_t = \left(\tilde{W}_{v_1,t}, \tilde{W}_{v_2,t}, \boldsymbol{r}_{t-1}^T \tilde{W}_{v_1,t}\right)^T, \tag{5.18}$$

where $\tilde{W}_{v_i,t} = W_{v_i,t} - \mu_{v_i}$ for $i = 1, 2$, and $\boldsymbol{r}_{t-1} = \big(f(P_{t-1}), \ldots, f(P_{t-h})\big)^T$.

Based on the theory in Giacomini & White (2006), we can construct the test statistic

$$S = (n-h)\, \overline{\boldsymbol{V}}^T\, \hat{\Sigma}_V^{-1}\, \overline{\boldsymbol{V}} \sim \chi^2_{h+2}\,, \tag{5.19}$$

where $\hat{\Sigma}_V$ is an estimator of $\mathrm{cov}(\boldsymbol{V}_t)$. To derive the expressions for this estimator it is convenient to rewrite (5.18) as

$$\boldsymbol{V}_t = \mathrm{diag}(\tilde{W}_{v_1,t}, \tilde{W}_{v_2,t}, \tilde{W}_{v_1,t}, \ldots, \tilde{W}_{v_1,t})\tilde{\boldsymbol{r}}_{t-1}\,, \tag{5.20}$$

where $\tilde{\boldsymbol{r}}_{t-1} = \big(1, 1, f(P_{t-1}), \ldots, f(P_{t-h})\big)^T$. We can then use a conditional expectation argument similar to that used to derive (5.12) to show that

$$\Sigma_V = \underbrace{\begin{pmatrix} \sigma^2_{g_1} & \sigma_{g_1,g_2} & \sigma^2_{g_1} & \ldots & \sigma^2_{g_1} \\ \sigma_{g_1,g_2} & \sigma^2_{g_2} & \sigma_{g_1,g_2} & \ldots & \sigma_{g_1,g_2} \\ \sigma^2_{g_1} & \sigma_{g_1,g_2} & \sigma^2_{g_1} & \ldots & \sigma^2_{g_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{g_1} & \sigma_{g_1,g_2} & \sigma^2_{g_1} & \ldots & \sigma^2_{g_1} \end{pmatrix}}_{A_{\boldsymbol{g}}} \circ \mathrm{E}\left(\tilde{\boldsymbol{r}}_{t-1}\, \tilde{\boldsymbol{r}}_{t-1}^T\right) \tag{5.21}$$

where $\circ$ denotes Hadamard product of two matrices. As before we can use the estimator

$$\hat{\Sigma}_V = (n-h)^{-1} A_{\boldsymbol{g}} \circ \sum_{t=h+1}^{n} \boldsymbol{r}_{t-1}\, \boldsymbol{r}_{t-1}^T\,, \tag{5.22}$$

where $A_{\boldsymbol{g}}$ denotes the first matrix in the Hadamard product in (5.21). Alternatively we can use the true value of Σ_V under the null hypothesis of iid uniform (P_t), with

$$\mathrm{E}\left(\tilde{\boldsymbol{r}}_{t-1}\, \tilde{\boldsymbol{r}}_{t-1}^T\right) = \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 0 & 0 & \sigma_r^2 & & \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & & & \sigma_r^2 \end{pmatrix} \tag{5.23}$$

where $\sigma_r^2 = \mathrm{E}\left(f(P_t)^2\right)$. It is convenient to set $f(P_t) = \tilde{W}_{v_1,t}$ so that $\sigma_r^2 = \sigma_{v_1}^2$.

As an alternative to (5.18), we have also considered the $(h+1)$-vector

$$\boldsymbol{V}_t = M_{t-1}(\tilde{W}_{v_1,t}, \tilde{W}_{v_2,t})^T\,, \tag{5.24}$$

where M_{t-1} is a $(h+1) \times 2$ matrix with columns $\boldsymbol{r}_{1,t-1}$ and $\boldsymbol{r}_{2,t-1}$, where $\boldsymbol{r}_{j,t-1} = \big(1, f_j(P_{t-1}), \ldots, f_j(P_{t-h})\big)^T$ for $j = 1, 2$. The is the conditional calibration test proposed by Nolde & Ziegel (2016). We find that this structure is less powerful in detecting departure from uniformity of P_t. To understand this, we can re-write (5.24) as

$$\boldsymbol{V}_t = \boldsymbol{r}_{1,t-1}\tilde{W}_{v_1,t} + \boldsymbol{r}_{2,t-1}\tilde{W}_{v_2,t}\,. \tag{5.25}$$

Hence, by considering (5.24), we are actually testing some linear combination of $\tilde{W}_{v_1,t}$ and $\tilde{W}_{v_2,t}$. The power of the test will be reduced since the correlation information between $\tilde{W}_{v_1,t}$ and $\tilde{W}_{v_2,t}$ is "diluted".

5.2.3 Conditional probitnormal score test

This fits into the framework of the conditional bispectral test. Recall from Proposition 3.5 that the score vector for the truncated probitnormal distribution in (3.73) can be written $\boldsymbol{S}_t(\boldsymbol{\theta}_0) = \big(S_{1,t}(\boldsymbol{\theta}_0), S_{2,t}(\boldsymbol{\theta}_0)\big)^T = \boldsymbol{W}_{\boldsymbol{v},t} - \boldsymbol{\mu}_{\boldsymbol{v}}$, almost surely, in terms of the weighted realized PIT values vector $\boldsymbol{W}_{\boldsymbol{v},t} = (W_{v_1,t}, W_{v_2,t})^T$ and a mean vector $\boldsymbol{\mu}_{\boldsymbol{v}} = (\mu_{v_1}, \mu_{v_2})^T$.

The theory of the previous section carries over with $\tilde{W}_{v_1,t} = S_{1,t}(\boldsymbol{\theta}_0)$ and $\tilde{W}_{v_2,t} = S_{2,t}(\boldsymbol{\theta}_0)$. We recall that the elements of the covariance matrix $\Sigma_{\boldsymbol{v}} = I(\boldsymbol{\theta}_0)$ of $\boldsymbol{W}_{\boldsymbol{v},t}$ are given in Appendix A.

5.2.4 Choice of factor for the conditional spectral and bispectral test

To test for an absence of serial correlation, it is sometimes more effective to consider transformation of the form $f(P_t) = f^*(\tilde{P}_t)$ for some function f^*, with $\tilde{P}_t = |2P_t - 1|$.

To understand this better, suppose we consider the series of losses (L_t) generated by a stationary GARCH process of the form $L_t = \sigma_t Z_t$, where (Z_t) are iid mean

zero and variance one residuals generated from some symmetrical distribution. In this case, the acf of (L_t) is zero. However, the acf of $(|L_t|)$ at lag $h > 0$ is positive, since by construction the acf of (σ_t) at lag $h > 0$ is positive. We now consider the uniformly distributed processes $U_t = F_t(L_t)$ and $\tilde{U}_t = \tilde{F}_t(|L_t|)$ for $t = 1, \ldots, n$, where $\tilde{F}_t$ denotes the distribution of $|L_t|$. By symmetry, $\tilde{F}(x) = |2F(x) - 1|$ so that $\tilde{U}_t = |2U_t - 1|$.

We will consider transformation of the form

$$\begin{aligned} f^*(\tilde{P}_t) &= \int_{\beta_1}^{\beta_2} I_{\{\tilde{P}_t > u\}} \mathrm{d}u \qquad (5.26) \\ &= \min(\beta_2, \max(\tilde{P}_t, \beta_1)) - \beta_1 \\ &= \begin{cases} 0 & \frac{1}{2}(1-\beta_1) \leqslant P_t \leqslant \frac{1}{2}(1+\beta_1)\,, \\ \tilde{P}_t - \beta_1 & \frac{1}{2}(1-\beta_2) < P_t < \frac{1}{2}(1-\beta_1) \ \text{ or } \ \frac{1}{2}(1+\beta_1) < P_t < \frac{1}{2}(1+\beta_2) \\ \beta_2 - \beta_1 & P_t \leqslant \frac{1}{2}(1-\beta_2) \ \text{ or } \ P_t \geqslant \frac{1}{2}(1+\beta_2) \end{cases} \\ &= 2\left(\int_{\frac{1}{2}(1+\beta_1)}^{\frac{1}{2}(1+\beta_2)} I_{\{1-P_t > u\}} \mathrm{d}u + \int_{\frac{1}{2}(1+\beta_1)}^{\frac{1}{2}(1+\beta_2)} I_{\{P_t > u\}} \mathrm{d}u \right). \qquad (5.27) \end{aligned}$$

The above transformation is the same as truncation $\tilde{P}_t$ to the range (β_1, β_2). Note that since the ACF of $f^*(\tilde{P}_t)$ remains the same for strictly increasing f^*, the ACF of $f^*(\tilde{P}_t)$ in (5.26) depends on β_1 and β_2 only. The above transformation is convenient since if P_t is uniform, $\tilde{P}_t$ will be uniform as well.

In the simulations studies later, we will focus on the continuous weighting case, where $W_{v,t}$ is defined in (3.33) as

$$W_{v,t} = \int_{\alpha_1}^{\alpha_2} g(u) I_{\{P_t > u\}} \mathrm{d}u\,. \qquad (5.28)$$

To simplify the formulas used for size correction, it is convenient to set β_1 and β_2 such that $\frac{1}{2}(1+\beta_1) \leqslant \alpha_1$ and $\frac{1}{2}(1+\beta_2) \geqslant \alpha_2$.

5.2.5 A different form for the conditional spectral and bispectral test

From the previous section, we note that $f^*(\tilde{P}_t)$ in (5.26) is more effective in detecting serial dependence. Hence, it would be sensible to tweak the conditional spectral and bispectral test in Section 5.2.1 and Section 5.2.2 to make better use of $f^*(\tilde{P}_t)$.

Recall that the untruncated component of $f^*(\tilde{P}_t)$ contains the information from the left tail $\left(\frac{1}{2}(1-\beta_2), \frac{1}{2}(1-\beta_1)\right)$ of P_t and the right tail $\left(\frac{1}{2}(1+\beta_1), \frac{1}{2}(1+\beta_2)\right)$ of P_t. For risk management purpose, when testing the hypothesis in (3.30), we are usually only interested in the evaluation of the right tail of the loss distribution. To minimize the deviation from our risk management objectives, we need to ensure that the test statistic is constructed in a way such that $f^*(\tilde{P}_t)$ is only used to test for absence of serial correlation. Due to the above reasoning, we will center $f^*(\tilde{P}_t)$ using its sample mean, which we denote by

$$\tilde{X}_t = f^*(\tilde{P}_t) - \hat{\mu}_X \,, \tag{5.29}$$

where $\hat{\mu}_X = \frac{1}{n}\sum_{t=1}^{n} f^*(\tilde{P}_t)$.

Conditional spectral test. As an alternative to testing the vector $\boldsymbol{V}_t = \boldsymbol{r}_{t-1}\tilde{W}_{v,t}$, with $\boldsymbol{r}_t = (1, f(P_{t-1}), \ldots, f(P_{t-h}))^T$ as proposed in Section 5.2.1, would be to consider the size $(h+1)$ vector

$$\boldsymbol{V}_t = \left(\tilde{W}_{v_1,t}, \boldsymbol{r}_{t-1}^T\tilde{X}_t\right)^T, \tag{5.30}$$

where $\boldsymbol{r}_{t-1} = (\tilde{X}_{t-1}, \ldots, \tilde{X}_{t-h})^T$. We set the covariance matrix to be

$$\hat{\Sigma}_V = \mathrm{diag}(\sigma_v^2, \hat{\sigma}_X^4, \ldots, \hat{\sigma}_X^4)\,, \tag{5.31}$$

where $\hat{\sigma}_X^2 = \frac{1}{n}\sum_{t=1}^{n}\tilde{X}_t^2$ is the sample variance of $\tilde{X}_t$. By doing so, we assume that $(\tilde{X}_t)$ are iid under the null, but make no assumptions on the distributional form of $(\tilde{X}_t)$. We then construct the test statistic

$$S = (n-h)\,\overline{\boldsymbol{V}}^T\,\hat{\Sigma}_V^{-1}\,\overline{\boldsymbol{V}} \sim \chi^2_{h+1}\,, \tag{5.32}$$

where $\overline{\boldsymbol{V}} = (n-h)^{-1}\sum_{t=h+1}^{n}\boldsymbol{V}_t$.

Conditional bispectral test. As an alternative to testing the vector

$$\boldsymbol{V}_t = (\tilde{W}_{v_1,t}, \tilde{W}_{v_2,t}, \boldsymbol{r}_{t-1}^T\tilde{W}_{v_1,t})^T \tag{5.33}$$

with $\boldsymbol{r}_t = \left(f(P_{t-1}), \ldots, f(P_{t-h})\right)^T$ as proposed in Section 5.2.2, would be to consider the size $(h+2)$ vector

$$\boldsymbol{V}_t = \left(\tilde{W}_{v_1,t}, \tilde{W}_{v_2,t}, \boldsymbol{r}_{t-1}^T\tilde{X}_t\right)^T, \tag{5.34}$$

where $\boldsymbol{r}_{t-1} = (\tilde{X}_{t-1}, \ldots, \tilde{X}_{t-h})^T$. We set the covariance matrix to be

$$\hat{\Sigma}_V = \begin{pmatrix} \sigma^2_{v_1} & \sigma_{v_1,v_2} & 0 & \ldots & 0 \\ \sigma_{v_1,v_2} & \sigma^2_{v_2} & 0 & \ldots & 0 \\ 0 & 0 & \hat{\sigma}^4_X & & \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & & & \hat{\sigma}^4_X \end{pmatrix}, \tag{5.35}$$

where $\hat{\sigma}^2_X = \frac{1}{n}\sum_{t=1}^{n} \tilde{X}_t^2$ is the sample variance of $\tilde{X}_t$. We then construct the test statistic

$$S = (n-h)\, \overline{\boldsymbol{V}}^T\, \hat{\Sigma}_V^{-1}\, \overline{\boldsymbol{V}} \sim \chi^2_{h+2}\,, \tag{5.36}$$

where $\overline{\boldsymbol{V}} = (n-h)^{-1} \sum_{t=h+1}^{n} \boldsymbol{V}_t$.

5.2.6 Size correction for the conditional spectral and bispectral test

Depending on the choice of factors that we use, the conditional spectral and bispectral test statistic S in (5.11) and (5.19) can have very poor size, especially when h is large, and size correction may be required. Similar to Nass test, we find c and v such that the first two moments of $c\,S$ matches the first two moments of the χ^2_v random variable, i.e.

$$c\,S \underset{H0}{\overset{d}{\sim}} \chi^2_\nu\,, \quad \text{with} \quad c = \frac{2\,\mathrm{E}\,(S)}{\mathrm{var}(S)} \text{ and } \nu = c\,\mathrm{E}\,(S)\,. \tag{5.37}$$

The theory used in Giacomini & White (2006) is based on matching the mean of the test statistic S with the chi-squared rv. Hence, for the conditional spectral test, the mean of the test statistic is given by $\mathrm{E}\,(S) = h+1$. Similarly, for the conditional bispectral test, the mean of the test statistic is given by $\mathrm{E}\,(S) = h+2$.

For the special case when factor $f(P_t) = \tilde{W}_{v,t}$ is used, the calculations for $\mathrm{var}(S)$ for the conditional spectral and bispectral tests are given in Appendix B.1. For the case when a generic factor $f(P_t) - \mu_f$ is used, where μ_f is the mean of $f(P_t)$ calculated under the assumption that (P_t) are iid uniform, the calculations for $\mathrm{var}(S)$ for the conditional spectral and bispectral tests are given in Appendix B.2. The calculations for $\mathrm{var}(S)$ for the conditional spectral and bispectral test when implemented based on suggestions in Section 5.2.5 are given in Appendix B.3.

5.3 Tests based on blocking

Another way to test for serial independence is by simply taking blocks of data, i.e. we divide the n realized PIT values into N_B blocks of size B. The motivation for doing so is that we have observed that the size performance of the conditional spectral and bispectral tests in Section 5.2.1 and Section 5.2.2 constructed using only $(W_{v_1,t})$ and $(W_{v_2,t})$ (i.e. we set the factors $f(P_t) = \tilde{W}_{v_1,t}$) to have rather poor size performance, especially when lag h is large. By taking larger blocks of data, we hope to improve the size performance at the cost of a lower power, without the need to resort to size correction techniques. For simplicity, we will use non-overlapping blocks.

5.3.1 Block spectral test

We define block sums and block products of weighted realized PIT values

$$\boldsymbol{Y}_{v,b} = \left(\sum_{t=s_b}^{bB} (W_{v,t} + k), \prod_{t=s_b}^{B} (W_{v,t} + k) \right)^T, \quad b = 1, \ldots, N_B, \quad s_b = (b-1)B+1, \tag{5.38}$$

for some constant $k > 0$ to ensure that the block product takes strictly positive values, which is necessary for the test to work. For simplicity, we set $k = \mu_B - \mu_v$ for some value $\mu_B > \mu_v$, and contruct the test based on μ_B. By doing so, we shift the expected block mean under the null hypothesis 3.30 from μ_v to μ_B.

Proposition 5.1. Let $\overline{\boldsymbol{Y}}_v = N_B^{-1} \sum_{b=1}^{N_B} \boldsymbol{Y}_{v,b}$. Under the null hypothesis 3.30, for fixed B,

$$\sqrt{N_B} \left(\overline{\boldsymbol{Y}}_v - \boldsymbol{\mu}_{\boldsymbol{Y}} \right) \tag{5.39}$$

is asymptotically $N_2(\boldsymbol{0}, \Sigma_{\boldsymbol{Y}})$ distributed, where $\boldsymbol{\mu}_{\boldsymbol{Y}} = (B\mu_B, \mu_B^B)^T$ and

$$\Sigma_{\boldsymbol{Y}} = \begin{pmatrix} B\sigma_v^2 & B\sigma_v^2 \mu_B^{B-1} \\ B\sigma_v^2 \mu_B^{B-1} & (\sigma_v^2 + \mu_B^2)^B - \mu_B^{2B} \end{pmatrix}. \tag{5.40}$$

Proof. We denote $\tilde{W}_{v,t} = W_{v,t} - \mu_v$. Under the null hypothesis 3.30 the vectors $\boldsymbol{Y}_{v,b} = (Y_{v,b,1}, Y_{v,b,2})^T$ are iid random vectors with mean $\boldsymbol{\mu_Y}$ and covariance matrix $\Sigma_{\boldsymbol{Y}}$. To calculate these moments, we observe that

$$\begin{aligned}
\mathrm{E}\left(Y_{v,b,1}Y_{v,b,2}\right) &= \mathrm{E}\left(\sum_{t=s_b}^{bB}(\tilde{W}_{v,t}+\mu_B)\prod_{t=s_b}^{bB}(\tilde{W}_{v,t}+\mu_B)\right) \\
&= \mathrm{E}\left(\sum_{t=s_b}^{bB}\left((\tilde{W}_{v,t}+\mu_B)^2\prod_{i\neq t}(\tilde{W}_{v,i}+\mu_B)\right)\right) \\
&= B\,\mathrm{E}\left((\tilde{W}_{v,1}+\mu_B)^2(\tilde{W}_{v,2}+\mu_B)\cdots(\tilde{W}_{v,B}+\mu_B)\right) \\
&= B\left(\sigma_v^2+\mu_B^2\right)\mu_B^{B-1}\,, \qquad (5.41)
\end{aligned}$$

and hence

$$\mathrm{cov}(Y_{v,b,1}, Y_{v,b,2}) = \mathrm{cov}\left(\sum_{t=s_b}^{bB}(\tilde{W}_{v,t}+\mu_B), \prod_{t=s_b}^{bB}(\tilde{W}_{v,t}+\mu_B)\right) = B\sigma_v^2\mu_B^{B-1}\,. \qquad (5.42)$$

Moreover

$$\begin{aligned}
\mathrm{var}\left(\prod_{t=s_b}^{bB}(\tilde{W}_{v,t}+\mu_B)\right) &= \mathrm{E}\left(\prod_{t=s_b}^{bB}(\tilde{W}_{v,t}+\mu_B)^2\right) - \left(\mathrm{E}\left(\prod_{t=s_b}^{bB}(\tilde{W}_{v,t}+\mu_B)\right)\right)^2 \\
&= \left(\sigma_v^2+\mu_B^2\right)^B - \mu_B^{2B}. \qquad (5.43)
\end{aligned}$$

The result is then simply an application of the central limit theorem in the multivariate case. □

Proposition 5.1 implies that for sufficiently large N_B the test statistic

$$S = N_B\left(\overline{\boldsymbol{Y}}_v - \boldsymbol{\mu_Y}\right)^T\Sigma_{\boldsymbol{Y}}^{-1}\left(\overline{\boldsymbol{Y}}_v - \boldsymbol{\mu_Y}\right) \sim \chi_2^2\,. \qquad (5.44)$$

Note that this test can also be applied to the case when we use the degenerate weight measure consisting of point mass at the single level α, which yields a test that we will refer to as the block binomial and the formulas for the required moments are simply $\mu_v = 1-\alpha$ and $\sigma_v^2 = \alpha(1-\alpha)$.

5.3.2 Block bispectral test

We can extend the blocking approach to the bispectral test by taking block product of one of the sequences of weighted realized PIT. For $b = 1, \ldots, N_B$ we form the

vectors

$$\boldsymbol{Y}_{\boldsymbol{v},b} = \left(\sum_{t=s_b}^{bB} (\tilde{W}_{v_1,t} + \mu_B), \sum_{t=s_b}^{bB} (\tilde{W}_{v_2,t} + \mu_B), \prod_{t=s_b}^{B} (\tilde{W}_{v_1,t} + \mu_B) \right)^T , \tag{5.45}$$

where $s_b = (b-1)B + 1$ and we require that $\mu_B > \mu_{v_1}$.

Proposition 5.2. Let $\overline{\boldsymbol{Y}}_{\boldsymbol{v}} = N_B^{-1} \sum_{b=1}^{N_B} \boldsymbol{Y}_{\boldsymbol{v},b}$. Under the null hypothesis 3.30, for fixed B,

$$\sqrt{N_B} \left(\overline{\boldsymbol{Y}}_{\boldsymbol{v}} - \boldsymbol{\mu}_{\boldsymbol{Y}} \right) \tag{5.46}$$

is asymptotically $N_3(\mathbf{0}, \Sigma_{\boldsymbol{Y}})$ distributed, where $\boldsymbol{\mu}_{\boldsymbol{Y}} = (B\mu_B, B\mu_B, \mu_B^B)^T$ and

$$\Sigma_{\boldsymbol{Y}} = \begin{pmatrix} B\sigma_{v_1}^2 & B\sigma_{v_1,v_2} & B\sigma_{v_1}^2 \mu_B^{B-1} \\ B\sigma_{v_1,v_2} & B\sigma_{v_2}^2 & B\sigma_{v_1,v_2} \mu_B^{B-1} \\ B\sigma_{v_1}^2 \mu_B^{B-1} & B\sigma_{v_1,v_2} \mu_B^{B-1} & (\sigma_{v_1}^2 + \mu_B^2)^B - \mu_B^{2B} \end{pmatrix} , \tag{5.47}$$

with $\sigma_{v_1,v_2} = \operatorname{cov}(W_{v_1,t}, W_{v_2,t}) = \mathrm{E}\left(W_{v_1,t} W_{v_2,t}\right) - \mu_{v_1}\mu_{v_2}$.

Proof. Once again this is a simple application of the multivariate CLT. Most of the calculations follow easily from those in the proof of Proposition 5.1. In addition, we use the fact that

$$\begin{aligned} &\mathrm{E}\left(\sum_{t=s_b}^{bB} (\tilde{W}_{v_2,t} + \mu_B) \prod_{t=s_b}^{bB} (\tilde{W}_{v_1,t} + \mu_B) \right) \\ &= \mathrm{E}\left(\sum_{t=s_b}^{bB} \left((\tilde{W}_{v_2,t} + \mu_B)(\tilde{W}_{v_1,t} + \mu_B) \prod_{i \neq t} (\tilde{W}_{v_1,t} + \mu_B) \right) \right) \\ &= B\mathrm{E}\left((\tilde{W}_{v_2,1} + \mu_B)(\tilde{W}_{v_1,1} + \mu_B)(\tilde{W}_{v_1,2} + \mu_B) \cdots (\tilde{W}_{v_1,B} + \mu_B) \right) \\ &= B\left(\sigma_{v_1,v_2} + \mu_B^2 \right) \mu_B^{B-1} , \end{aligned} \tag{5.48}$$

from which it follows that

$$\operatorname{cov}\left(\sum_{t=s_b}^{bB} (\tilde{W}_{v_2,t} + \mu_B), \prod_{t=s_b}^{bB} (\tilde{W}_{v_1,t} + \mu_B) \right) = B\sigma_{v_1,v_2} \mu_B^{B-1} . \tag{5.49}$$

□

Proposition 5.2 implies that for sufficiently large N_B the test statistic

$$N_B \left(\overline{\boldsymbol{Y}}_{\boldsymbol{v}} - \boldsymbol{\mu}_{\boldsymbol{Y}}\right)^T \Sigma_{\boldsymbol{Y}}^{-1} \left(\overline{\boldsymbol{Y}}_{\boldsymbol{v}} - \boldsymbol{\mu}_{\boldsymbol{Y}}\right) \sim \chi_3^2 . \tag{5.50}$$

Note that (5.45) is one of the few possible choices of vectors that we can consider. In particular, we have also considered the symmetric vector

$$\boldsymbol{Y}_{\boldsymbol{v},b} = \left(\prod_{t=s_b}^{B} (\tilde{W}_{v_1,t} + \mu_B), \prod_{t=s_b}^{B} (\tilde{W}_{v_2,t} + \mu_B) \right) . \tag{5.51}$$

We find that in this case, while the size performance of the test is better, it comes at the cost of worst power performance, possibly due to the loss of information from the block sum. Additionally, we have also considered the symmetric vector

$$\boldsymbol{Y}_{\boldsymbol{v},b} = \left(\sum_{t=s_b}^{bB} (\tilde{W}_{v_1,t} + \mu_B), \sum_{t=s_b}^{bB} (\tilde{W}_{v_2,t} + \mu_B), \prod_{t=s_b}^{B} (\tilde{W}_{v_1,t} + \mu_B), \prod_{t=s_b}^{B} (\tilde{W}_{v_2,t} + \mu_B) \right) . \tag{5.52}$$

For this case, we find that both size and power performance of the test are slightly worst when compared to testing the vector in (5.45), possibly due to the additional information gain from the second block product is unable to compensate for the increase in the degrees of freedom of the test.

5.3.3 Block probitnormal score test

This fits into the framework of the previous section. Recall that the score vector for the truncated probitnormal distribution in (3.73) can be written $\boldsymbol{S}_t(\boldsymbol{\theta}_0) = (S_{1,t}(\boldsymbol{\theta}_0), S_{2,t}(\boldsymbol{\theta}_0))^T = \boldsymbol{W}_{\boldsymbol{v},t} - \boldsymbol{\mu}_{\boldsymbol{v}}$, almost surely, in terms of a vector of weighted realized PIT values $\boldsymbol{W}_{\boldsymbol{v},t} = (W_{v_1,t}, W_{v_2,t})^T$ and a mean vector $\boldsymbol{\mu}_{\boldsymbol{v}} = (\mu_{v_1}, \mu_{v_2})^T$, where v_1 and v_2 is given by (3.80) and (3.81). The theory of the previous section carries over with $\tilde{W}_{v_1,t} = S_{1,t}(\boldsymbol{\theta}_0)$ and $\tilde{W}_{v_2,t} = S_{2,t}(\boldsymbol{\theta}_0)$.

For $b = 1, \ldots, N_B$ we form the vectors

$$\boldsymbol{Y}_{\boldsymbol{v},b} = \left(\sum_{t=s_b}^{bB} (S_{1,t}(\boldsymbol{\theta}_0) + \mu_B), \sum_{t=s_b}^{bB} (S_{2,t}(\boldsymbol{\theta}_0) + \mu_B), \prod_{t=s_b}^{B} (S_{1,t(\boldsymbol{\theta}_0)} + \mu_B) \right)^T , \tag{5.53}$$

where $s_b = (b-1)B+1$ and we require that $\mu_B > \mu_{v_1}$. We then apply Proposition 5.2 and the fact that $\text{cov}(\boldsymbol{W}_{\boldsymbol{v},t}) = \text{cov}(\boldsymbol{S}_t(\boldsymbol{\theta}_0)) = I(\boldsymbol{\theta}_0)$ to infer that

$$\sqrt{N_B} \left(\overline{\boldsymbol{Y}}_{\boldsymbol{v}} - \boldsymbol{\mu}_{\boldsymbol{Y}}\right) \tag{5.54}$$

is asymptotically $N_3(\mathbf{0}, \Sigma_{\boldsymbol{Y}})$ distributed, where $\boldsymbol{\mu_Y} = (B\mu_B, B\mu_B, \mu_B^B)^T$ and $\Sigma_{\boldsymbol{Y}}$ is given by (5.47) with $\sigma_{v_1}^2 = I(\boldsymbol{\theta}_0)_{1,1}$, $\sigma_{v_2}^2 = I(\boldsymbol{\theta}_0)_{2,2}$ and $\sigma_{v_1,v_2} = I(\boldsymbol{\theta}_0)_{1,2}$. The necessary formulas for the Fisher information matrix are given in Appendix A.

Chapter 6 Simulation studies: Explicit tests of independence

6.1 Size of tests

In this section, we will try to understand the size of the tests described in Chapter 5 when applied to iid uniform data. We will focus on the tests based on martingale difference (MD) property in Section 5.2 and tests based on blocking in Section 5.3. We have omit the portmanteau tests based on acf in Section 5.1 as we found that they are similar to the MD tests. For ease of analysis, we have colour coded the tables, where green indicates good results ($\leqslant 6\%$ for the size; $\geqslant 70\%$ for the power); red indicates poor results ($\geqslant 9\%$ for the size; $\leqslant 30\%$ for the power); dark red indicates very poor results ($\geqslant 12\%$ for the size; $\leqslant 10\%$ for the power).

6.1.1 Size of tests based on martingale difference property

We carry forward the tests B99, SP.U, SP.L, SP.E.200, SP.UL, SP.UE.200 and PNS from Chapter 4. For all of these tests we implement the corresponding martingale difference versions. We used the following factors to test for serial independence:

Factor W We have used the lagged, centred exception indicators $f(P_t) = \tilde{W}_{v,t}$. This choice reflects our goal of constructing a backtest purely based on the transformed realized PIT values ($\tilde{W}_{v,t}$). This is implemented based on methods described in Section 5.2, where we evaluate the covariance matrix Σ_V in (5.13) and (5.21) under the null hypothesis (3.30).

Factor WX.Full Here, we set $f(P_t) = f^*(\tilde{P}_t)$, where f^* in defined in (5.26), and $\tilde{P}_t = |2P_t - 1|$. The motivation for using this is described in Section 5.2.4. We set $\beta_1 = 0.0005$ and $\beta_2 = 0.9995$ to be consistent with our usual convention of truncating extreme realized PIT values.

Factor X This is the version that make use of $(\tilde{X}_t)$ as described in Section 5.2.5, where $\tilde{X}_t$ is defined by (5.29). We consider the truncation parameters ($\beta_1 = 2\alpha_1 - 1, \beta_2 = 2\alpha_2 - 1$), ($\beta_1 = 0.5, \beta_2 = 0.9995$) and ($\beta_1 = 0.0005, \beta_2 = 0.9995$),

which we denote the correspoding tests as Factor X, Factor X.Half and Factor X.Full. By expanding the range (i.e. by reducing the amount of truncation), we expect the size and power of the tests to improve due to the increase in information contained by the factors.

Table 13 shows the estimated size of the above tests based on 10,000 replications, where the realized PIT values is truncated to the levels $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$. At lag $h = 1$, most tests exhibit good size performance. As h increases, the size of Factor W becomes very poor. The size of Factor WX.Full and Factor X is better than those of Factor W, but is still over-sized for small sample size ($n = 250$). This is because (P_t) and $(\tilde{P}_t)$ are truncated to the right extreme tail, which leads to slower convergence rate of the test statistics to their asymptotic distribution. Factor X.Half and Factor X.Full with less truncation have much better size.

Table 14 shows the results when size corrections based on Section 5.2.6 are performed. We see that for Factor W, Factor WX.Full and Factor X, the size of the conditional spectral and bispectral tests are now much better. We do notice that there is a tendency for the tests to be under-sized, especially for Factor X when h is large. This is partly because when performing size correction, we evaluate all expectations that involve $\tilde{X}_t$ using (B.54). The size of the conditional binomial test still exhibits poor size for some lags, but the overall size performance is much better compared to Table 13.

h	n	factor | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	PNS
1	250	W	3.7	6.2	4.7	3.7	7.0	7.0	7.1
		WX.Full	5.2	5.0	5.3	6.3	5.9	6.1	6.1
		X	6.0	5.6	5.8	6.1	6.3	6.3	6.6
		X.Half	4.5	4.5	4.4	4.7	5.3	5.3	5.5
		X.Full	5.0	5.1	4.9	5.1	5.3	5.4	5.7
	500	W	5.8	7.1	5.8	4.5	7.5	7.6	7.1
		WX.Full	4.6	4.8	4.8	5.3	5.4	5.5	5.5
		X	5.9	5.2	5.2	5.2	5.5	5.8	6.0
		X.Half	4.6	4.8	4.8	5.0	5.2	5.3	5.4
		X.Full	4.8	4.9	4.8	4.9	5.2	5.3	5.4
4	250	W	8.9	12.2	8.7	5.4	12.0	12.1	10.2
		WX.Full	8.7	7.6	8.7	10.6	8.4	8.6	7.6
		X	8.7	7.2	7.3	7.4	7.4	7.4	7.3
		X.Half	4.9	4.9	5.0	4.9	5.1	5.1	5.3
		X.Full	5.4	5.3	5.4	5.4	5.4	5.7	5.5
	500	W	17.2	12.5	11.0	7.3	11.9	12.1	9.8
		WX.Full	6.8	6.4	7.0	8.4	6.8	7.0	6.5
		X	8.2	6.8	6.7	6.9	6.8	6.8	6.9
		X.Half	5.1	4.9	5.0	5.1	5.3	5.4	5.5
		X.Full	5.2	5.0	5.1	5.2	5.5	5.5	5.6
9	250	W	17.4	17.4	13.5	8.3	16.6	16.7	14.5
		WX.Full	12.8	11.2	13.1	15.7	11.5	11.8	10.0
		X	9.1	7.7	7.7	7.8	7.7	7.7	7.8
		X.Half	5.1	5.0	5.2	5.3	5.3	5.3	5.5
		X.Full	5.2	5.1	5.2	5.4	5.2	5.3	5.4
	500	W	32.0	16.0	16.3	11.2	15.5	15.4	12.9
		WX.Full	10.2	8.6	10.1	12.1	8.8	9.2	7.8
		X	10.2	7.4	7.5	7.5	7.5	7.6	7.6
		X.Half	5.3	5.2	5.2	5.3	5.5	5.6	5.5
		X.Full	5.0	4.8	4.9	5.1	5.0	5.2	5.2

Table 13: Estimated size of the two-sided conditional spectral and bispectral tests based on martingale difference property. Results are based on 10000 replications, with $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$.

h	n	factor | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	PNS
1	250	W	2.7	3.3	2.4	3.8	3.0	3.2	3.2
		WX.Full	4.0	4.0	4.1	4.1	4.4	4.4	4.5
		X	3.0	3.0	3.1	3.1	3.2	3.2	3.1
		X.Half	4.0	4.1	3.9	4.0	4.3	4.1	4.1
		X.Full	4.5	4.6	4.4	4.3	4.5	4.3	4.4
	500	W	4.8	3.7	3.1	2.5	3.7	3.9	4.2
		WX.Full	3.9	4.2	4.0	3.8	4.6	4.6	4.6
		X	3.5	3.3	3.4	3.4	3.8	3.8	3.8
		X.Half	4.4	4.6	4.5	4.4	4.7	4.7	4.7
		X.Full	4.6	4.7	4.5	4.4	4.7	4.6	4.8
4	250	W	8.8	4.6	4.2	3.5	4.3	4.3	4.4
		WX.Full	4.6	4.6	4.6	4.6	4.7	4.6	4.8
		X	2.7	1.9	1.9	1.8	1.8	1.8	1.8
		X.Half	4.1	4.2	4.2	4.1	4.1	4.0	3.9
		X.Full	4.8	5.0	4.9	4.5	4.8	4.5	4.4
	500	W	2.1	3.5	4.4	3.5	3.6	3.6	4.2
		WX.Full	4.7	4.6	4.4	4.5	4.7	4.8	4.8
		X	2.7	2.6	2.7	2.7	2.7	2.7	2.7
		X.Half	4.7	4.4	4.5	4.6	4.7	4.6	4.8
		X.Full	4.9	4.8	4.8	4.7	5.1	5.0	4.9
9	250	W	2.8	4.3	5.0	4.1	4.2	4.3	4.4
		WX.Full	4.8	5.2	4.8	4.7	5.1	5.1	5.2
		X	1.8	1.0	1.0	1.0	1.0	1.0	1.1
		X.Half	4.2	4.1	4.2	4.2	4.0	4.1	4.0
		X.Full	4.7	4.4	4.4	4.4	4.3	4.4	4.3
	500	W	3.1	4.1	4.7	4.6	4.1	4.1	4.7
		WX.Full	5.1	5.0	5.1	5.0	5.1	5.1	5.1
		X	2.1	2.0	2.0	2.0	2.0	2.0	2.0
		X.Half	4.8	4.8	4.7	4.7	4.9	4.9	4.8
		X.Full	4.6	4.5	4.6	4.8	4.6	4.8	4.6

Table 14: Estimated size of the two-sided conditional spectral and bispectral tests based on martingale difference property, with size correction performed. Results are based on 10000 replications, with $\alpha_1 = 0.975$ *and* $\alpha_2 = 0.9995$.

6.1.2 Size of tests based on blocking

We carry forward the tests B99, SP.U, SP.L, SP.E.200, SP.UL, SP.UE.200 and PNS from Chapter 4. For all of these tests we implement the blocking method based on Section 5.3. We set μ_B such that the correlation between the block sums and block product of weighted realized PIT values is close to one, as we found that this method produces the best size performance.

Table 15 shows the estimated size of the tests based on blocking, where the realized PIT values is truncated to the levels $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$. We used 10,000 replications. On the contrary to the tests based on MD property, we observed that the size improves as the block size B increases. For the block bispectral tests, when sample size is small ($n = 250, 500$), we require a large block size ($B = 10$) for the size to be acceptable. In general, the size performance of the block tests is much better than their conditional tests counterparts (without size correction).

n	B | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	PNS
250	2	2.5	4.9	3.7	3.1	5.9	6.0	9.7
	3	3.7	6.0	4.6	3.8	7.0	7.0	6.8
	5	5.9	6.4	5.5	4.5	7.1	7.3	6.5
	10	5.5	4.9	6.3	5.3	6.2	6.5	5.9
500	2	3.6	6.2	4.5	3.5	6.9	7.0	7.3
	3	5.6	6.8	5.5	4.3	7.3	7.5	6.8
	5	10.0	5.6	6.3	5.0	6.4	6.7	6.3
	10	4.2	5.2	5.7	5.6	5.8	5.9	5.3
1000	2	6.4	7.2	6.0	4.3	7.4	7.5	6.7
	3	10.6	6.3	6.4	4.8	6.5	6.8	5.7
	5	4.8	5.0	5.9	5.6	5.6	5.7	5.7
	10	4.1	5.4	5.3	5.7	5.5	5.8	4.7
2000	2	11.2	6.5	6.6	4.8	6.2	6.3	5.8
	3	4.2	5.3	5.8	5.5	5.4	5.3	5.4
	5	5.9	5.1	5.3	5.9	5.0	5.0	5.3
	10	5.4	5.1	5.4	5.6	5.2	5.2	4.3

Table 15: Estimated size of the two-sided block spectral and bispectral tests. Results are based on 10000 replications, with $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$.

6.2 Experiment one: ARMA process

6.2.1 Experimental design

As mentioned in Section 4.4.1, the historical simulation (HS) and filtered historical simulation (FHS) models are widely used by banks. Figure 7 shows the acf plots when HS and FHS methods with parameter $\lambda = 0.94$ are applied to S&P 500 returns from January 2010 to December 2015. The top row shows the acf of (P_t), and the bottom row shows the acf of $(\tilde{P}_t)$, where $\tilde{P}_t = |2P_t - 1|$. We observed that while the acf of (P_t) looks fine, the acf of $(\tilde{P}_t)$ is not, especially for the HS model, where we observed persistent large acf across many lags, whereas for the FHS model, even though the acf decays quickly, it is large for lag less than 5.

For the simulation study, we will attempt to replicate this form of misspecification. We do this by generating a sequence (Z_t) from a Gaussian ARMA model with mean zero and variance one and transform the sequence to have a standard uniform distribution by taking $\tilde{U}_t = \Phi(Z_t)$. We consider the

ARMA process (Z_t) is an ARMA(1,1) process, with $Z_t = \varphi Z_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$, where $\epsilon_t, \epsilon_{t-1}, \ldots$ are white noise error terms. We set the parameters $\varphi = 0.95$ and $\theta = -0.85$. The chosen parameter values correspond closely to the values of parameter estimates obtained when an ARMA(1,1) model is fitted to the probit-transformed realized PIT values of the HS model in Figure 7.

AR process (Z_t) is an AR(1) process, with $Z_t = \varphi Z_{t-1} + \epsilon_t$, where ϵ_t is white noise. We set the autoregressive parameter $\varphi = 0.5$. We chose this process to understand the power of the tests when acf decays quickly.

We then form a further uniform sequence (U_t) by setting

$$U_t = \tfrac{1}{2}(1 + \tilde{U}_t)^{B_t}(1 - \tilde{U}_t)^{(1-B_t)}\,, \tag{6.1}$$

where (B_t) is a series of iid Bernoulli variables with mean 0.5. Note that the uniform sequences (U_t) and $(\tilde{U}_t)$ are related by $\tilde{U}_t = |2U_t - 1|$. In particular (U_t) mimics the realized PIT values obtained when the distribution is correctly estimated but serial dependence coming from the stochastic volatility is neglected. Figure 8 shows the

acf plots of (U_t) and $(\tilde{U}_t)$ when (Z_t) follows the ARMA and AR processes.

We then construct the test data (P_t) by setting $P_t = \Phi(F^{-1}(U_t))$ where F is the distribution function of normal, Student-$t5$ or Student-$t3$, standardized to have zero mean and unit variance. This construction is very similar to Section 4.1.1, where the forecast model is the standard normal distribution. In the case when F is normal, (P_t) is a serially correlated sequence of uniformly distributed data. In the other cases, (P_t) are serially correlated and non-uniform, which correspond to the cases when the forecast models neglect the modeling of stochastic volatility as well as consistently underestimate the tail of the loss distribution.

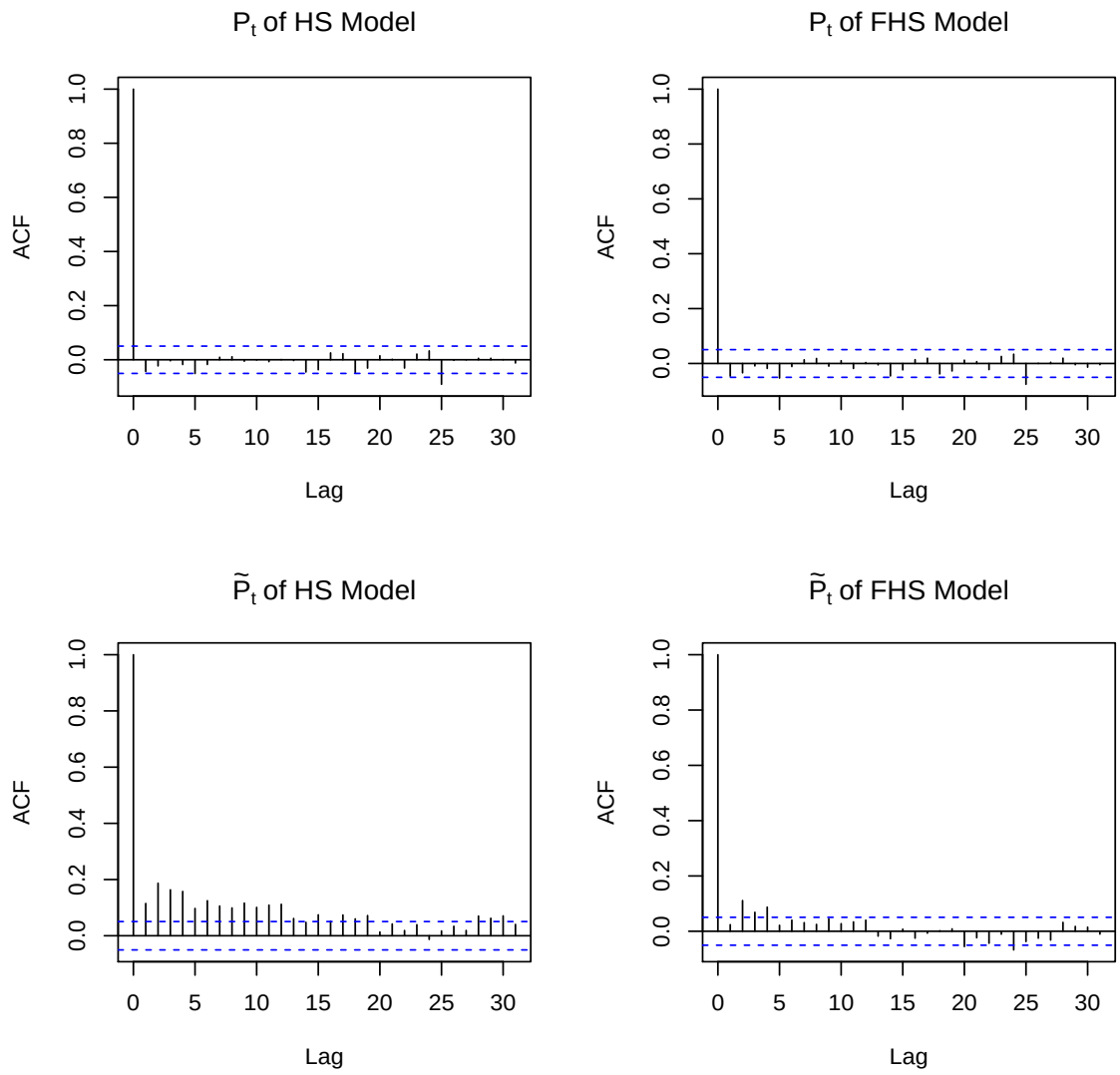

Figure 7: Acf plots of realized PIT values (P_t) and transformed realized PIT ($\tilde{P}_t$) obtained from applying historical simulation (HS) and filtered historical simulation (FHS) to S&P 500 returns from January 2010 to December 2015.

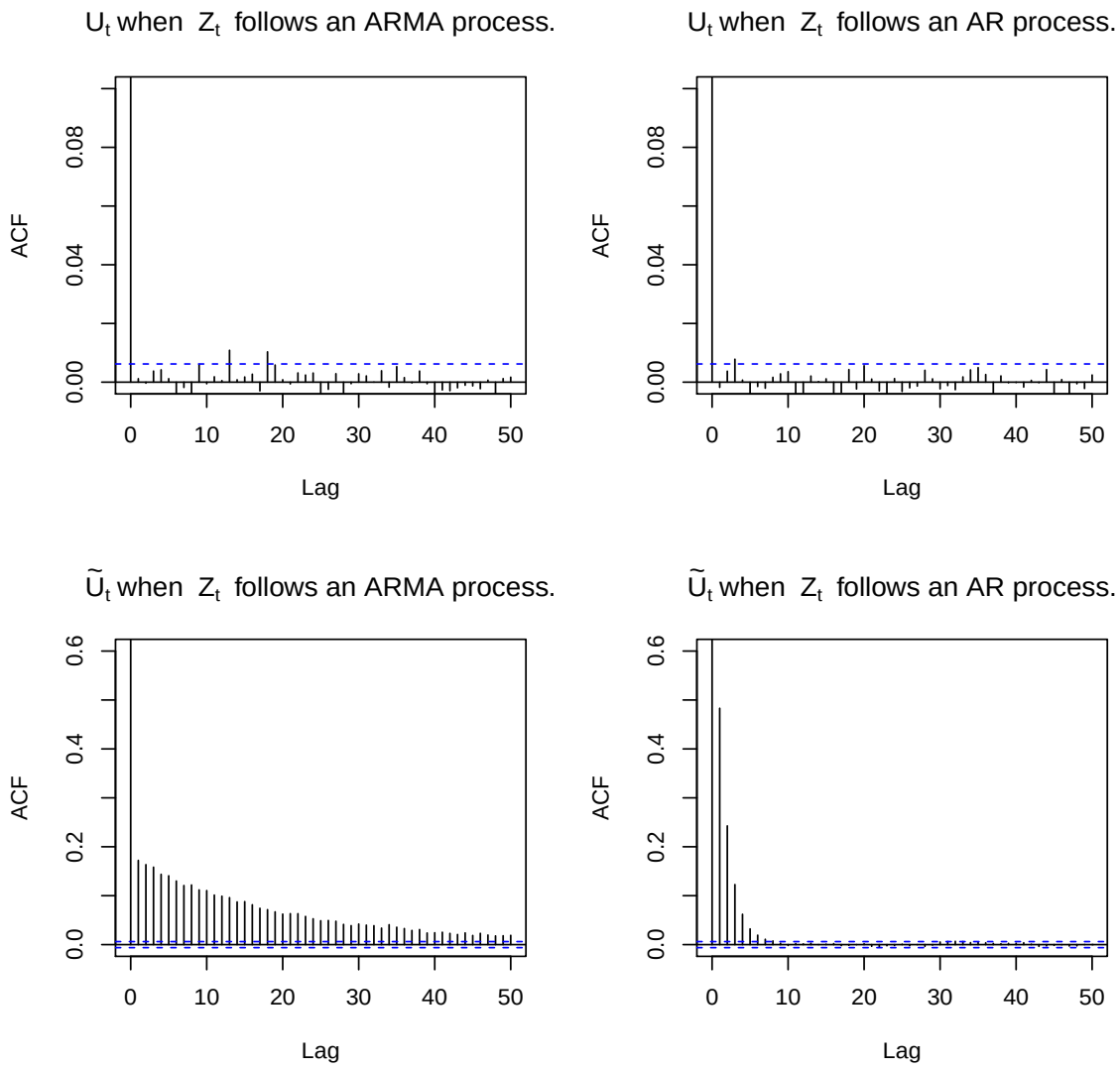

Figure 8: acf plots of of (U_t) *and* $(\tilde{U}_t)$ *when* (Z_t) *is an ARMA(1,1) process with parameters* $\varphi = 0.95$ *and* $\theta = -0.85$*, and when* (Z_t) *is an AR(1) process with autoregressive parameter* $\varphi = 0.5$*.*

6.2.2 Power of tests in the case when F is normal

We first consider the case when F is normal, and see how the power of the tests vary when taking different lags. We consider taking lag $h = 1, 4, 9$, with corresponding block size $B = 2, 5, 10$. 'ITT' denotes the test type, with 'None' refering to the standard spectral and bispectral test, 'Block' refers to block spectral and bispectral test, implemented based on Section 6.1.2, and 'MD' are the martingale difference test, implemented based on Section 6.1.1 with size correction.

ARMA process. First, we look at the case when (Z_t) is an ARMA(1,1) process with parameters $\varphi = 0.95$ and $\theta = -0.85$. In this case, the acf is moderately sized, but persistent. The results is shown in Table 16.

First, we compare the Block test and the MD.W test. This is a fair comparison, and both tests are constructed using only $(\tilde{W}_{v_1,t})$ and $(\tilde{W}_{v_2,t})$. At first glance, it would appear that Block tests are more powerful than MD.W tests, however, recall from Section 6.1.1 and Section 6.1.2 that due to size correction, MD.W tests are mostly slightly undersized, whereas the Block tests in many cases are slightly oversized. Hence, it would be fair to say that both test types are actually similar in terms of power performance.

Next, we notice that for the spectral tests, the Block, MD.W and MD.WX.Full tests becomes less powerful as the weighting function shifts from uniform weight to exponential weight. To understand this, we refer to Figure 9, which plots the acf of (W_t), where $W_t = I_{\{P_t > \alpha\}}$, at levels $\alpha = 0.975$, $\alpha = 0.99$ and $\alpha = 0.9995$, when (Z_t) is an ARMA(1,1) process with parameters $\varphi = 0.95$ and $\theta = -0.85$. We see that the acf of W_t becomes smaller as the level α increases. Hence, the weighted realized PIT values $(\tilde{W}_{v,t})$ which places more weight at higher levels becomes less effective in detecting misspecification in dynamics.

As we would have expect, using $(\tilde{X}_t)$ instead of $(\tilde{W}_{v_1,t})$ to test for serial independence will result in a more powerful test. This is observed by comparing the power of MD.W with MD.X, since they share the same amount of truncation. Also, as we decrease the amount of truncation, the tests becomes more powerful. This is

observed by comparing MD.X with MD.X.Half and MD.X.Full, where we observe MD.X.Full to be most powerful This is because we gain more information regarding the misspecification in dynamics as we reduce the amount of truncation. To see this, we refer to Figure 10, which plots the acf of $(\tilde{X}_t)$ at various truncation levels, when (Z_t) is an ARMA(1,1) process with parameters $\varphi = 0.95$ and $\theta = -0.85$.

We now focus on MD.WX.Full, MD.X.Half and MD.X.Full, since from Table 14, we know that the size is good for all h and all n that we are considering. We observe that as the lag increases, the power of the test increases. From Figure 8, we know that the acf is roughly the same for lag 1 to lag 9. As we increase the lag of the tests, we are essentially placing more weight (degree of freedom) on the component that test for serial independence, and less weight (degree of freedom) on the component that tests for departure from uniformity.

Finally, since (P_t) is uniformly distributed, for the bispectral tests, changing the weighting function of $(\tilde{W}_{v_2,t})$ does not affect the power of the tests.

AR process. We now look at the case when (Z_t) is an AR(1) process with autoregressive parameter $\varphi = 0.5$. In this case, the acf is very large at lag 1, but decays very quickly. The results is shown in Table 17.

Some of the observations here are similar to those from Table 16. First, the block tests and MD.W tests are similar in power. Next, for the MD tests, MD.X is more powerful than MD.W, and reducing the amount of truncation further increases power. For the spectral tests, the Block tests, MD.W tests and MD.WX.Full tests becomes increasingly less powerful as the weighting shifts from uniform weight to exponential weight. Finally, as we changes the weight functions of $(\tilde{W}_{v_2,t})$ in the bispectral tests, the power remains roughly the same.

The key difference between the result in Table 17 and Table 16 is that now, as we increase the lag, the power decreases. This is because by construction, the test statistic places equal weight for misspecification for all lags. We try to understand this in an intuitive and non-rigorous way. Suppose we denote the acf at lag k by $\rho(k)$, and we are testing up to lag h. The MD test statistic tests the departure from uniformity with weight $\frac{2}{h+2}$, and tests for serial independence based on the average

acf $\bar{\rho}_h = \frac{1}{h}\sum_{k=1}^{h}\rho(k)$ with weight $\frac{h}{h+2}$. Since the acf decays quickly, the average acf $\bar{\rho}_h$ decreases as h increases, and hence we observe a decrease in power.

ARMA process versus AR process. When we compare the results of the standard spectral and bispectral test in Table 16 and Table 17, we notice that rejection rate for the ARMA process is higher than that of the AR process. As we have mentioned at the beginning of Chapter 5, the standard spectral test and bispectral tests implicitly test for serial independence, because when we calculate the variance under the null hypothesis (3.30), we set $\mathrm{var}(\sum_{t=1}^{n} W_{v,t}) = \sum_{t=1}^{n}\mathrm{var}(W_{v,t})$, which assumes that $(W_{v,t})$ are iid. To understand this better, we refer to Gordy et al. (2017). They have shown that in the binomial case with $W_t = I_{\{P_t>\alpha\}}$, where $\alpha > 0.5$ and (P_t) is the process defined by

$$P_t = \Phi\Big(F^{-1}\Big(\tfrac{1}{2}(1+\Phi(Z_t))^{B_t}(1-\Phi(Z_t))^{(1-B_t)}\Big)\Big) \tag{6.2}$$

for an independent Bernoulli process (B_t) with success probability 0.5 and a Gaussian ARMA process (Z_t) with mean zero and variance one, the acf $\rho_W(k)$ of (W_t) is related to the acf $\rho_Z(k)$ of (Z_t) by

$$\rho_W(k) = \frac{C^{\mathrm{Ga}}_{\rho_Z(k)}(2-2\tilde{\alpha}, 2-2\tilde{\alpha}) - 4(1-\tilde{\alpha})^2}{\tilde{\alpha}(1-\tilde{\alpha})}\,, \tag{6.3}$$

where $\tilde{\alpha} = F(\Phi^{-1}(\alpha))$ and C^{Ga}_{ρ} denotes the bivariate Gaussian copula with parameter ρ.

We can then calculate

$$\lim_{n\to\infty} n\,\mathrm{var}\left(\overline{W}_n\right) = \alpha(1-\alpha)\left(1 + 2\sum_{k=1}^{\infty}\rho_W(k)\right)\,, \tag{6.4}$$

where $\overline{W}_n = \frac{1}{n}\sum_{t=1}^{n}(W_t - (1-\alpha))$. We then estimate numerically $\lim_{n\to\infty}\mathrm{var}(\sqrt{n}\overline{W}_n) = ((1+\eta)\sigma_W)^2$ for the ARMA and AR process in the case of the B99 test using (6.3) and (6.4). For the AR(1) model we obtain $\eta \approx 1.11$ whereas for the ARMA(1,1) model we obtain $\eta \approx 1.23$. It is the larger value in the latter which leads to the slightly higher rejection rate for the B99 test in the ARMA(1,1) experiment.

6.2.3 Power of tests in the case when F is normal, $t5$ and $t3$

We now consider the case when F is normal, $t5$ and $t3$. We consider lag $h = 4$ (and $B = 5$), since from previous observations, h should not be too large or too small. The results is shown in Table 18 and Table 19. Much of the conclusions remains the same as those in Table 16 and Table 17 at lag 4.

As we vary F from normal to $t5$ to $t3$, the conclusions are the same as we have previously seen in Table 7 in Section 4.1.5, where spectral tests which places more weight in the tail for more powerful, and generally bispectral tests are more powerful than spectral tests.

The key difference for the results in Table 18 and Table 19 when compared to Table 16 and Table 17 is that now, in the case when F is $t5$ and $t3$, MD.WX.Full is more powerful than MD.X, since $(\tilde{W}_{v_1,t})$ is more effective than $(\tilde{X}_t)$ at detecting departure from uniformity of P_t.

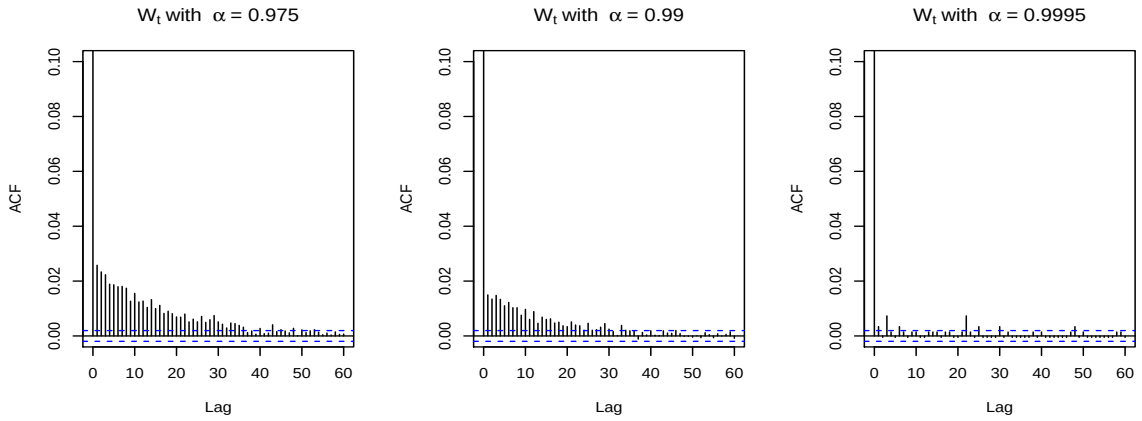

Figure 9: Acf plots of $(W_t) = (I_{\{P_t > \alpha\}})$ *at various levels, when* (Z_t) *is an ARMA(1,1) process with parameters* $\varphi = 0.95$ *and* $\theta = -0.85$.

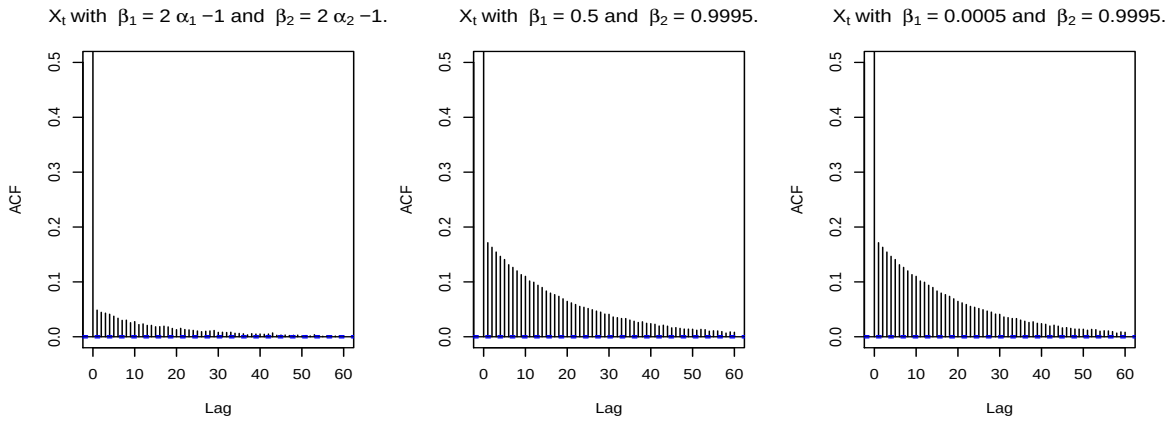

Figure 10: Acf plots of $(\tilde{X}_t)$ *at various truncation levels, when* (Z_t) *is an ARMA(1,1) process with parameters* $\varphi = 0.95$ *and* $\theta = -0.85$.

6.3 Experiment two: standard GARCH process

6.3.1 Experimental design

Here the backtesting set-up is similar to that used in Section 4.2.1, except that the experiment is conducted in a time-series setup. The true data-generating mechanism for the losses is a stationary GARCH model with Student-t innovations, which combines the features of stochastic volatility and heavy tails, since these features are usually observed in the market risk-factor changes and trading book losses.

We will simulate losses from a GARCH(1,1) model with Student-t innovations. The model for simulated losses (L_t) takes the form

$$L_t = \sigma_t Z_t, \quad \sigma_t^2 = \varphi_0 + \varphi_1 L_{t-1}^2 + \theta_1 \sigma_{t-1}^2 \,, \tag{6.5}$$

where (Z_t) is an iid Student-t distribution with ν degrees of freedom, standardized to have zero mean and unit variance. The parameters have been chosen by fitting this model to the (negative) S&P index log-returns for the period January 2007 to December 2012. We have chosen this particular period for model fitting as it includes the 2008 financial crisis as well as the relatively stable period after. The parameters of the GARCH equation are $\varphi_0 = 2.16 \times 10^{-6}$, $\varphi_1 = 0.112$ and $\theta_1 = 0.887$ while the degrees of freedom of the Student-t innovation distribution is $\nu = 5.46$. Figure 11 shows the estimated parameters when we fit the GARCH.t model to the S&P 500 losses from January 2016 minus x-years to December 2015. The blue dotted line shows the chosen period of estimation and corresponding parameters.

We will assume that the bank uses a rolling window of n_2=250 and n_2=500 days to calibrate its trading book models, where re-calibration takes place every 10 trading days. We will consider backtest with sample sizes n=250 and n=500 days, since this is the typical amount of data available to the regulators from banks. We then repeat the experiment 1000 times to estimate rejection rates for each forecaster.

We consider the following forecasting methods:

Oracle: the forecaster knows the underlying model as well as the exact parameter values.

h	n	ITT | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	PNS
1	250	None	8.7	11.2	9.0	7.9	10.0	9.9	11.1
		Block	8.0	11.8	10.8	8.2	12.6	12.2	17.9
		MD.W	8.3	10.1	8.0	9.4	9.8	10.2	10.8
		MD.WX.Full	12.5	13.7	13.0	11.0	13.5	13.0	14.5
		MD.X	10.8	14.2	14.1	13.5	13.4	13.3	13.5
		MD.X.Half	40.1	40.7	40.3	38.4	35.5	33.6	34.4
		MD.X.Full	56.6	58.3	57.0	54.8	49.8	49.0	48.1
	500	None	8.5	13.1	11.7	8.9	10.3	10.3	11.6
		Block	9.5	15.3	12.4	9.9	14.9	15.5	16.9
		MD.W	10.7	11.7	9.2	7.4	11.8	12.4	13.2
		MD.WX.Full	13.8	18.7	15.8	13.4	17.1	16.3	18.5
		MD.X	15.6	19.3	18.6	18.2	17.4	17.5	18.3
		MD.X.Half	68.4	70.4	68.9	68.1	64.9	64.1	64.1
		MD.X.Full	83.1	83.3	83.0	82.6	79.1	78.9	78.5
4	250	None	8.7	11.2	9.0	7.9	10.0	9.9	11.1
		Block	13.4	15.1	13.8	11.0	16.0	15.7	15.2
		MD.W	17.1	12.8	11.0	9.7	12.2	12.2	13.8
		MD.WX.Full	13.6	17.3	14.7	12.4	17.3	17.1	19.7
		MD.X	11.3	14.0	13.5	13.1	13.6	13.8	13.4
		MD.X.Half	59.5	59.3	59.4	57.6	58.1	56.9	57.6
		MD.X.Full	74.3	74.3	73.7	73.5	73.1	71.9	72.5
	500	None	8.5	13.1	11.7	8.9	10.3	10.3	11.6
		Block	19.4	16.3	15.1	12.0	16.4	16.6	17.2
		MD.W	9.5	14.7	14.1	11.0	15.1	14.9	17.1
		MD.WX.Full	18.1	26.3	21.1	16.9	24.8	24.2	30.1
		MD.X	16.7	23.6	23.1	23.1	23.2	23.4	23.8
		MD.X.Half	88.9	89.3	89.2	88.9	87.8	87.3	87.5
		MD.X.Full	95.7	96.1	95.9	95.8	95.3	95.0	94.9
9	250	None	8.7	11.2	9.0	7.9	10.0	9.9	11.1
		Block	12.1	14.3	15.2	11.9	15.7	15.0	16.4
		MD.W	9.7	13.9	12.4	10.3	14.0	14.1	14.4
		MD.WX.Full	15.3	18.9	15.4	12.8	18.3	18.3	21.1
		MD.X	7.9	12.5	11.9	11.7	12.5	12.3	12.1
		MD.X.Half	61.5	61.5	61.1	60.7	60.8	60.2	60.7
		MD.X.Full	75.1	75.5	75.3	75.0	75.4	75.0	75.2
	500	None	8.5	13.1	11.7	8.9	10.3	10.3	11.6
		Block	10.9	17.4	16.8	14.6	17.9	17.8	18.7
		MD.W	12.1	16.5	14.0	11.1	16.3	16.4	19.0
		MD.WX.Full	19.1	28.3	23.1	18.7	27.7	27.6	33.6
		MD.X	16.5	25.2	25.0	24.6	25.0	24.8	25.2
		MD.X.Half	90.7	91.3	91.1	90.6	91.1	90.8	90.6
		MD.X.Full	96.6	96.6	96.6	96.6	96.4	96.3	96.4

Table 16: Estimated power of the two-sided standard, Block and MD spectral and bispectral tests, with size correction performed, when (Z_t) is an ARMA(1,1) process with parameters $\varphi = 0.95$ and $\theta = -0.85$. We vary the lag parameter h. Results are based on 10000 replications, with $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$.

h	n	ITT | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	PNS
1	250	None	6.9	8.5	6.2	6.9	7.7	7.9	8.4
		Block	12.5	20.1	16.2	12.6	21.4	21.0	29.8
		MD.W	18.0	23.7	18.6	16.1	21.9	22.0	22.3
		MD.WX.Full	25.8	38.1	31.6	23.2	34.4	33.4	39.2
		MD.X	31.9	50.8	50.9	50.8	49.7	49.2	48.2
		MD.X.Half	100.0	100.0	100.0	100.0	99.9	99.9	99.9
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	500	None	6.8	9.2	8.5	7.1	7.5	8.2	8.2
		Block	20.1	31.1	23.9	18.2	29.6	29.0	31.6
		MD.W	32.8	34.8	27.3	21.1	32.3	32.2	36.7
		MD.WX.Full	42.7	69.9	54.5	40.8	60.9	60.3	73.6
		MD.X	51.5	77.1	77.0	76.9	74.6	74.6	74.2
		MD.X.Half	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0
4	250	None	6.9	8.5	6.2	6.9	7.7	7.9	8.4
		Block	18.7	22.1	18.5	13.8	22.1	22.0	22.8
		MD.W	23.3	20.8	17.8	14.5	20.5	20.5	22.5
		MD.WX.Full	21.6	32.1	25.3	19.1	31.1	30.4	37.7
		MD.X	26.4	39.7	39.6	39.4	39.0	38.1	37.9
		MD.X.Half	99.9	99.9	99.9	99.9	99.9	99.8	99.8
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	500	None	6.8	9.2	8.5	7.1	7.5	8.2	8.2
		Block	34.6	31.1	27.4	20.9	29.7	30.7	30.8
		MD.W	16.3	28.6	25.7	19.9	28.2	28.0	33.1
		MD.WX.Full	37.4	56.6	45.3	33.2	52.8	52.4	65.1
		MD.X	46.7	69.3	69.0	69.2	68.0	68.1	67.6
		MD.X.Half	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0
9	250	None	6.9	8.5	6.2	6.9	7.7	7.9	8.4
		Block	12.4	17.8	17.5	13.3	18.8	18.7	20.0
		MD.W	9.4	14.9	14.5	12.6	15.1	15.3	18.2
		MD.WX.Full	17.8	24.8	20.4	15.1	23.7	23.7	29.7
		MD.X	17.8	28.6	28.5	28.2	28.1	27.7	28.0
		MD.X.Half	99.7	99.8	99.7	99.7	99.8	99.8	99.8
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	500	None	6.8	9.2	8.5	7.1	7.5	8.2	8.2
		Block	16.6	25.1	24.1	19.6	24.5	25.3	25.6
		MD.W	15.2	23.2	20.0	16.7	23.0	22.8	26.8
		MD.WX.Full	29.9	44.0	35.0	26.4	42.7	42.6	52.6
		MD.X	35.9	58.6	58.3	58.1	57.1	57.0	57.5
		MD.X.Half	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 17: Estimated power of the two-sided standard, Block and MD spectral and bispectral tests, with size correction performed, when (Z_t) is an AR(1) process with autoregressive parameter $\varphi = 0.5$. We vary the lag parameter h. Results are based on 10000 replications, with $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$.

F	n	ITT | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	PNS
Normal	250	None	8.7	11.2	9.0	7.9	10.0	9.9	11.1
		Block	13.4	15.1	13.8	11.0	16.0	15.7	15.2
		MD.W	17.1	12.8	11.0	9.7	12.2	12.2	13.8
		MD.WX.Full	13.6	17.3	14.7	12.4	17.3	17.1	19.7
		MD.X	11.3	14.0	13.5	13.1	13.6	13.8	13.4
		MD.X.Half	59.5	59.3	59.4	57.6	58.1	56.9	57.6
		MD.X.Full	74.3	74.3	73.7	73.5	73.1	71.9	72.5
	500	None	8.5	13.1	11.7	8.9	10.3	10.3	11.6
		Block	19.4	16.3	15.1	12.0	16.4	16.6	17.2
		MD.W	9.5	14.7	14.1	11.0	15.1	14.9	17.1
		MD.WX.Full	18.1	26.3	21.1	16.9	24.8	24.2	30.1
		MD.X	16.7	23.6	23.1	23.1	23.2	23.4	23.8
		MD.X.Half	88.9	89.3	89.2	88.9	87.8	87.3	87.5
		MD.X.Full	95.7	96.1	95.9	95.8	95.3	95.0	94.9
t5	250	None	22.5	22.9	26.2	33.9	32.8	36.7	41.5
		Block	24.4	27.9	29.2	32.6	34.4	38.1	41.2
		MD.W	28.4	25.0	24.7	26.0	26.8	27.6	32.3
		MD.WX.Full	24.7	28.9	32.7	37.1	33.7	36.4	40.9
		MD.X	16.4	18.8	22.1	26.5	25.5	28.4	31.0
		MD.X.Half	57.5	57.2	58.7	61.0	60.9	62.4	64.0
		MD.X.Full	75.3	76.4	76.5	77.6	77.2	77.8	79.6
	500	None	26.2	25.8	36.3	46.2	46.4	52.9	58.7
		Block	39.1	34.3	41.6	46.8	48.4	54.2	57.6
		MD.W	23.4	33.8	36.8	34.7	39.0	42.2	47.2
		MD.WX.Full	35.4	42.9	46.1	51.6	53.0	55.5	62.7
		MD.X	26.2	32.3	37.5	42.6	43.2	46.5	51.6
		MD.X.Half	86.3	86.4	87.1	88.4	89.6	90.1	91.5
		MD.X.Full	96.6	96.5	96.7	96.7	96.8	97.2	96.9
t3	250	None	18.6	19.0	23.9	34.2	36.0	42.6	52.1
		Block	21.5	23.0	26.4	33.0	35.5	42.1	46.2
		MD.W	25.4	23.0	22.8	25.4	25.3	27.8	33.0
		MD.WX.Full	18.9	23.7	28.8	37.2	34.2	39.3	47.6
		MD.X	15.3	15.8	19.3	24.9	25.6	30.2	34.3
		MD.X.Half	49.9	49.0	51.1	55.6	57.5	59.5	63.3
		MD.X.Full	73.0	74.1	75.7	77.0	77.8	79.0	82.3
	500	None	20.1	20.0	31.4	46.2	60.2	66.8	76.0
		Block	33.7	27.9	36.7	47.1	57.6	64.5	71.2
		MD.W	19.5	31.1	35.1	35.4	40.4	45.1	54.2
		MD.WX.Full	26.1	35.4	39.2	47.9	59.2	63.5	77.9
		MD.X	23.6	26.3	31.2	39.6	45.4	52.0	59.0
		MD.X.Half	78.2	77.8	79.9	81.9	87.5	89.4	92.0
		MD.X.Full	95.4	95.6	96.0	96.3	97.4	97.7	98.2

Table 18: Estimated power of the two-sided standard, Block and MD spectral and bispectral tests, with size correction performed, when (Z_t) is an ARMA(1,1) process with parameters $\varphi = 0.95$ and $\theta = -0.85$. We vary the choice of F. Results are based on 10000 replications, with $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$.

F	n	ITT | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	PNS
Normal	250	None	6.9	8.5	6.2	6.9	7.7	7.9	8.4
		Block	18.7	22.1	18.5	13.8	22.1	22.0	22.8
		MD.W	23.3	20.8	17.8	14.5	20.5	20.5	22.5
		MD.WX.Full	21.6	32.1	25.3	19.1	31.1	30.4	37.7
		MD.X	26.4	39.7	39.6	39.4	39.0	38.1	37.9
		MD.X.Half	99.9	99.9	99.9	99.9	99.9	99.8	99.8
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	500	None	6.8	9.2	8.5	7.1	7.5	8.2	8.2
		Block	34.6	31.1	27.4	20.9	29.7	30.7	30.8
		MD.W	16.3	28.6	25.7	19.9	28.2	28.0	33.1
		MD.WX.Full	37.4	56.6	45.3	33.2	52.8	52.4	65.1
		MD.X	46.7	69.3	69.0	69.2	68.0	68.1	67.6
		MD.X.Half	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0
t5	250	None	19.4	19.5	25.4	33.3	30.4	37.3	42.6
		Block	32.3	34.8	36.3	37.7	41.0	44.5	47.7
		MD.W	38.7	35.9	34.9	34.1	35.8	36.7	38.6
		MD.WX.Full	40.6	49.5	50.4	51.3	51.0	52.6	57.4
		MD.X	36.7	44.5	45.7	46.8	47.2	48.8	50.9
		MD.X.Half	99.6	99.5	99.5	99.5	99.5	99.5	99.6
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	500	None	25.7	24.3	35.3	46.1	45.2	52.3	59.9
		Block	55.5	48.9	52.7	54.4	57.5	62.5	65.6
		MD.W	36.1	49.4	49.9	47.2	51.9	53.7	59.1
		MD.WX.Full	63.8	80.2	77.9	76.2	81.5	82.9	87.5
		MD.X	61.7	75.2	76.0	77.6	78.4	80.0	81.1
		MD.X.Half	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0
t3	250	None	15.9	14.5	22.6	34.0	37.4	46.1	53.3
		Block	29.8	31.0	32.8	36.7	41.7	48.8	53.1
		MD.W	35.5	33.1	32.2	32.7	33.7	35.2	39.4
		MD.WX.Full	34.9	45.0	46.2	50.2	51.5	53.8	59.6
		MD.X	34.2	39.8	40.9	43.1	43.9	46.2	50.0
		MD.X.Half	97.9	97.9	98.0	97.7	97.9	98.3	98.2
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	500	None	19.9	17.6	29.2	46.0	60.1	67.3	77.3
		Block	50.0	44.3	48.4	53.2	63.2	68.9	76.4
		MD.W	30.8	45.2	48.7	47.2	51.4	54.2	61.9
		MD.WX.Full	57.9	77.0	74.7	74.2	84.3	86.2	91.4
		MD.X	58.7	66.8	68.5	71.7	75.6	77.4	80.1
		MD.X.Half	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		MD.X.Full	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 19: Estimated power of the two-sided standard, Block and MD spectral and bispectral tests, with size correction performed, when (Z_t) is an AR(1) process with autoregressive parameter $\varphi = 0.5$. We vary the choice of F. Results are based on 10000 replications, with $\alpha_1 = 0.975$ and $\alpha_2 = 0.9995$.

GARCH.t: the forecaster assumes the correct model structure, i.e. the GARCH(1,1) model with Student-t innovations, but is required to estimate the GARCH parameters as well as the degrees of freedom of the innovations.

GARCH.EVT: the forecaster uses a GARCH(1,1) model to estimate the dynamics of the losses, and then applies an EVT tail model to the residuals to estimate the innovation distribution and hence the realized PIT values of the conditional loss distribution. We used the Weissman (1978) model with additional regression-based smoothing to estimate the tail distribution, where parameter estimation procedures are based on those described in Gomes & Martins (2002), Fraga et al. (2003) and Gomes & Pestana (2007).

GARCH.norm: the forecaster assumes that losses follow a GARCH(1,1) model with standard normal innovation distribution.

ARCH.t: the forecaster assumes that losses follow a ARCH(1) model with Student-t innovations. In this case, the forecaster misspecifies the dynamics of the losses, but correctly guesses that the distribution of the innovations.

ARCH.norm: similar to GARCH.norm, except that the forecaster misspecifies the dynamics to be ARCH(1).

EWMA.HS: the forecaster uses the Exponential Weighted Moving Average (EWMA) model with $\lambda = 0.94$ to estimate the dynamics of the losses, and uses the linearly interpolated empirical function in (4.21) to estimate the innovation distribution and hence the realized PIT values of the conditional loss distribution. This method is often referred to as the filtered historical simulation (FHS) in practice.

HS: the forecaster applies the linearly interpolated empirical function in (4.21) to the data to estimate the realized PIT values. As well as completely neglecting the dynamics of losses, this method is to prone underestimating the tail of the loss distribution when the calibration window size n_2 is small.

For more details of the above methodologies, see McNeil et al. (2015), Chapter 9.

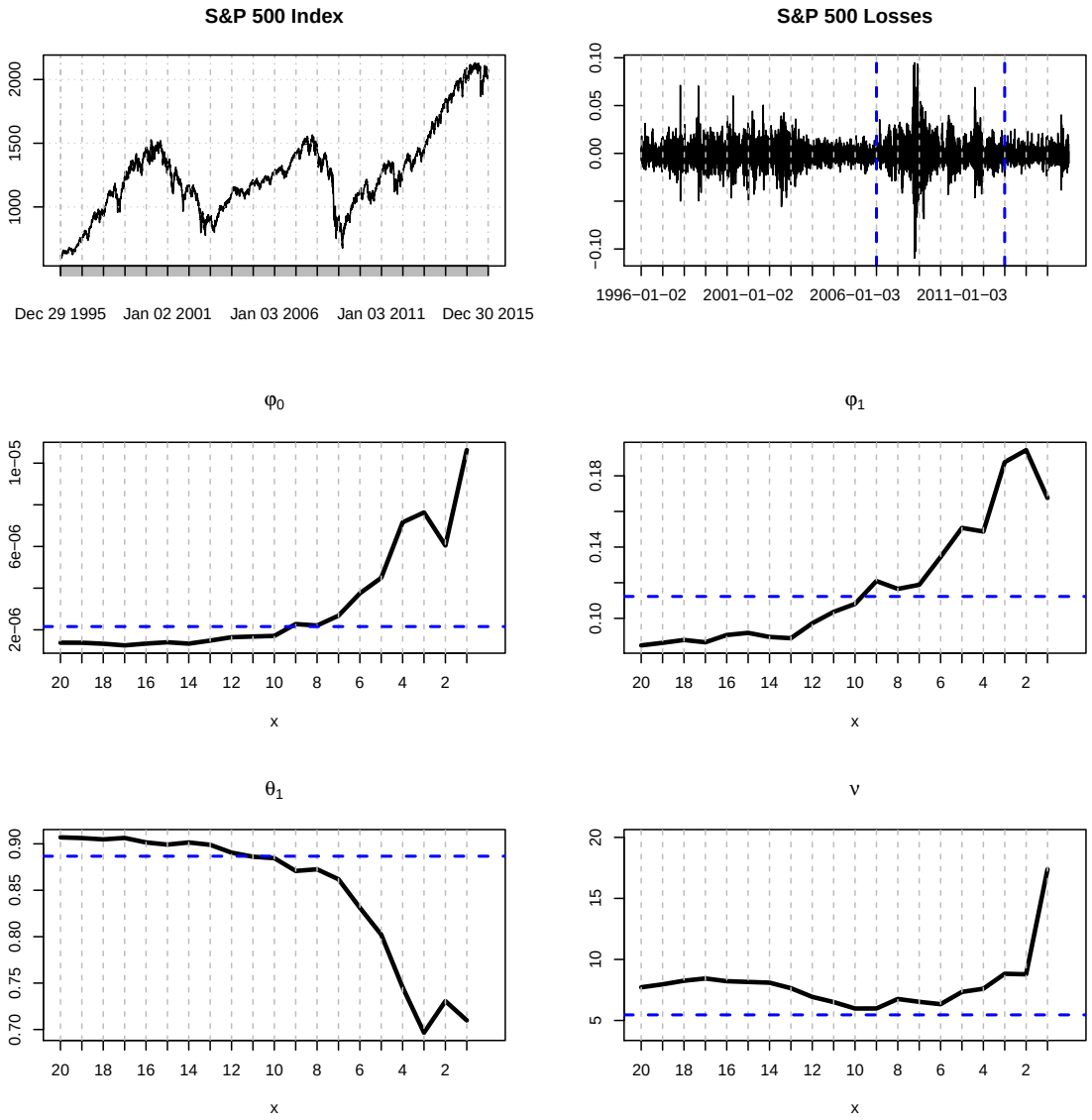

Figure 11: Estimated parameters when we fit the GARCH.t model to the S&P 500 losses from January (2016-x) to December 2015. The blue dotted line shows the chosen period of estimation and parameters.

6.4 Test results

We now look at the results in Table 20, Table 21 and Table 22. The structure of the table is similar to those in Section 6.2.2, except that the data generating process and forecast models are as described in Section 6.3.1, and forecast models now contain parameter estimation error. We have omitted the MD.X.Half tests as we find that the results is similar to the MD.Full tests.

Rejection rate of the good models. We will refer to the GARCH.t and GARCH.EVT models as the good models. The results are summarized in Table 20. Although the above models are 'good' only if the parameter estimation window n_2 is large enough, it is unclear how large n_2 should be. We have decided that $n_2 = 250$ is good enough, and used the size colouring in the table for ease of analysis.

Note that we have omitted the results when the forecast model $\widehat{F}_t$ is the 'Oracle', since we have already studied the size performance of the tests in Section 6.1.1 and Section 6.1.2.

When $n_2 = 250$, we observe that both models are rejected fairly frequently. The Block tests in particular seems penalize the parameter estimation error more heavily compared to the standard tests, followed by MD.W, MD.WX.Full and MD.X. In contrary, MD.X.Full is not very powerful in detecting parameter estimation error, especially in the case when $\widehat{F}_t$ is GARCH.t.

When $n_2 = 500$, the rejection rates for GARCH.t and GARCH.EVT are much smaller, especially for the MD.X and MD.X.Full tests.

Rejection rate of the bad models. We will refer to the GARCH.norm, ARCH.t and ARCH.norm models as the bad models. The results are summarized in Table 21, where we have used the power colouring in the table.

When $\widehat{F}_t$ is the GARCH.norm model, we observe that the standard tests and block tests are most powerful , followed by MD.WX.Full and MD.X.Full. It seems by reducing the amount of truncation on $(\tilde{X}_t)$, the tests are able to capture the departure

from uniformity of (P_t) better. This observation is made by comparing MD.X with MD.X.Full. Similar to observations from Section 4.1.5, bispectral tests are in general more powerful than spectral tests with PNS being the most powerful. When $\widehat{F}_t$ is the ARCH.t model, the observations are similar to those in Section 6.2.2, where MD.X.Full is most powerful, followed by MD.X and MD.WX.Full. We have omitted the results when $\widehat{F}_t$ is the ARCH.norm model, as we found that it is rather uninformative. We observed that in this case all tests have similar power performance where the rejection rates lie approximately in the range 90% to 100%.

Rejection rate of the industrial models. We will refer to the EWMA.HS and HS models as the industrial models. The results are summarized in Table 22. We have used the power colouring in the table, but this decision is arbitrary.

First, we focus on the case when $\widehat{F}_t$ is the EWMA.HS model. When $n_2 = 250$, the model has large estimation error in the tail. Hence, we would expect tests which are powerful in detecting the GARCH.norm misspecification in Table 21, namely the standard tests, Block tests and MD.WX.Full tests, to be powerful here as well, which is indeed the case. When $n_2 = 500$, this model have slight misspecification in both dynamics and innovation distribution, where the level of misspecification is roughly the same. It seems that the Block tests are slightly more powerful than the other tests in this case.

We now focus on the case when $\widehat{F}_t$ is the HS model. When $n_2 = 500$, most of the misspecification is with the dynamics of the test. Hence, we would expect tests which are powerful in detecting the ARCH.t misspecification in Table 21, namely the MD.X tests and MD.WX.Full tests, to be powerful here as well, which is indeed the case. When $n_2 = 250$, in addition to neglecting the dynamics of the simulated data, this model also has large estimation error in the tail. The misspecification in dynamics seems to dominates the misspecification in innovation distribution, resulting in MD.X.Full being the most powerful, followed by MD.X.

Conclusion. It is difficult to draw conclusions from Table 20 and Table 21 as to which tests should be preferred. In terms of the independence test type, Block tests

should be preferred when the goal is to detect departure from uniformity of the realized PIT values, followed by MD.WX.Full and MD.X.Full. If we wish to test for serial independence of the realized PIT values, MD.X.Full is most powerful, followed by MD.X and MD.WX.Full. For a well rounded test, it would seem that MD.X.Full is a good candidate, however, it has problems in detecting parameter estimation error. If this is a concern, then we should use MD.WX.Full. In terms of choice of weighting, while PNS seems to be most powerful, in practice, the choice of the weight functions depends on the purpose of the tests. For example, in a regulatory setting, motivated by testing the VaR at the 99% level, a humped weight function centered around the 99% level for the spectral tests may be preferred.

n_2	$\widehat{F}$	n	ITT | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	PNS
250	GARCH.t	250	None	17.0	17.2	16.7	15.3	13.5	12.3	15.7
			Block	15.7	15.6	15.5	10.1	15.7	12.1	17.6
			MD.W	14.4	12.2	6.7	6.7	12.2	12.2	13.9
			MD.WX.Full	20.4	15.0	17.1	15.7	17.1	17.0	15.4
			MD.X	11.5	11.8	11.8	11.8	11.8	11.7	11.7
			MD.X.Full	15.2	13.3	13.4	11.8	13.3	13.3	11.6
		500	None	13.8	14.1	15.9	16.0	14.1	14.6	16.1
			Block	29.4	27.1	23.4	16.2	23.6	25.7	27.3
			MD.W	15.7	19.5	18.0	14.2	17.9	17.8	23.1
			MD.WX.Full	17.5	13.8	25.0	23.4	17.5	17.8	18.0
			MD.X	19.0	19.3	19.4	21.3	21.3	19.5	17.7
			MD.X.Full	8.4	6.5	8.5	12.3	10.5	10.4	10.6
	GARCH.EVT	250	None	15.1	12.0	11.4	7.7	13.7	13.8	9.9
			Block	12.3	18.0	14.1	4.8	14.6	16.5	16.0
			MD.W	14.5	10.3	6.7	2.7	10.4	10.4	14.1
			MD.WX.Full	20.1	11.5	11.2	9.7	11.5	11.5	11.8
			MD.X	6.4	11.7	10.2	8.3	9.9	11.8	10.0
			MD.X.Full	15.5	15.5	13.7	10.2	13.7	13.5	8.1
		500	None	13.6	23.2	15.7	13.5	15.5	15.7	15.1
			Block	25.9	28.8	21.7	10.6	23.3	21.6	19.8
			MD.W	10.2	19.7	12.4	3.0	19.7	19.7	20.9
			MD.WX.Full	23.2	28.6	22.7	13.6	26.5	26.3	17.4
			MD.X	8.3	17.5	15.6	15.6	15.8	17.4	15.5
			MD.X.Full	18.8	15.5	15.3	11.9	15.3	16.9	15.1
500	GARCH.t	250	None	7.9	6.9	7.1	6.5	6.7	7.9	7.4
			Block	9.0	9.0	8.9	6.9	9.4	10.4	9.9
			MD.W	12.4	7.2	6.9	5.6	6.9	7.1	6.6
			MD.WX.Full	6.7	8.2	7.8	7.2	8.1	8.5	7.9
			MD.X	3.3	3.2	3.3	2.9	2.9	3.1	3.0
			MD.X.Full	5.6	4.8	4.7	5.6	5.2	5.8	5.8
		500	None	6.3	7.2	7.4	7.7	6.9	7.5	7.8
			Block	13.5	8.5	9.0	8.0	8.4	9.1	9.5
			MD.W	3.4	6.1	7.0	6.0	6.3	6.6	6.9
			MD.WX.Full	6.6	6.7	7.6	8.9	7.1	7.1	8.5
			MD.X	4.2	4.8	4.9	4.8	4.9	4.5	4.5
			MD.X.Full	5.3	5.1	6.1	6.3	5.6	5.7	6.4
	GARCH.EVT	250	None	10.1	8.8	8.3	7.2	8.8	9.8	8.5
			Block	11.5	12.6	10.8	7.2	13.6	13.3	11.5
			MD.W	15.3	8.7	7.2	5.9	8.1	8.2	8.8
			MD.WX.Full	9.4	10.5	9.6	6.0	9.1	9.5	8.2
			MD.X	4.7	3.5	3.4	3.2	3.7	3.6	3.5
			MD.X.Full	7.8	7.1	7.2	6.6	7.7	6.7	5.6
		500	None	8.5	8.7	8.9	7.6	7.8	7.6	5.9
			Block	16.1	10.9	11.9	8.3	10.2	10.0	9.2
			MD.W	4.7	8.6	8.5	6.5	7.7	7.8	6.2
			MD.WX.Full	8.4	9.4	8.6	6.5	8.1	8.3	7.6
			MD.X	4.8	4.3	4.1	3.7	4.2	4.1	3.9
			MD.X.Full	6.6	6.3	6.2	4.7	5.9	5.3	5.0

Table 20: Rejection rates for various good realized PIT estimation methods and various tests in the dynamic backtesting experiment. The data-generating mechanism for the losses is the GARCH.t process. Models are refitted after 10 days. Results are based on 1000 replications, with $\alpha_1 = 0.975$ *and* $\alpha_2 = 0.9995$.

n_2	$\widehat{F}$	n	ITT | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	PNS
250	GARCH.norm	250	None	25.3	21.9	31.2	37.3	39.0	40.1	47.4
			Block	22.0	25.8	29.4	36.8	37.3	41.6	45.7
			MD.W	26.2	23.7	23.6	23.5	22.0	22.1	29.2
			MD.WX.Full	30.7	26.8	33.2	39.2	29.4	35.4	39.3
			MD.X	15.2	19.3	17.9	20.1	17.9	22.0	24.0
			MD.X.Full	18.3	18.2	20.5	24.5	22.7	30.1	34.2
		500	None	43.0	40.7	50.7	57.4	53.4	63.2	68.7
			Block	44.9	46.7	47.8	57.0	53.4	66.9	70.4
			MD.W	27.0	35.3	37.4	32.2	39.4	45.1	56.5
			MD.WX.Full	44.0	45.8	58.1	64.4	54.9	64.3	62.7
			MD.X	30.4	27.0	31.3	41.2	37.7	43.3	47.2
			MD.X.Full	36.4	29.1	40.9	49.3	49.2	58.6	60.9
	ARCH.t	250	None	25.2	33.3	27.0	28.7	27.2	27.4	34.8
			Block	25.7	33.7	29.6	27.4	33.6	33.4	35.0
			MD.W	29.4	29.2	25.5	25.0	29.1	29.1	34.8
			MD.WX.Full	27.5	36.5	31.2	25.9	31.3	29.5	33.6
			MD.X	30.1	39.7	39.7	37.9	39.6	39.6	41.4
			MD.X.Full	49.7	53.1	53.2	53.1	53.1	51.2	49.5
		500	None	29.1	34.8	34.5	38.5	33.0	34.9	35.1
			Block	41.4	39.9	40.9	29.7	43.1	41.3	48.2
			MD.W	35.3	43.3	35.9	31.6	43.3	43.3	50.3
			MD.WX.Full	47.2	51.0	54.5	49.0	55.0	56.7	53.3
			MD.X	46.7	65.5	69.1	67.4	68.9	63.6	61.9
			MD.X.Full	82.4	84.2	85.8	85.7	82.2	82.3	82.0
500	GARCH.norm	250	None	20.5	19.7	25.1	33.6	27.4	33.4	38.0
			Block	17.8	22.6	25.2	29.9	29.7	34.7	37.1
			MD.W	21.1	17.4	16.8	18.4	17.8	19.5	24.6
			MD.WX.Full	18.2	20.4	26.0	32.1	25.2	28.9	31.4
			MD.X	5.2	6.1	8.7	12.1	10.8	13.9	17.8
			MD.X.Full	12.8	13.0	16.8	23.4	19.2	25.0	30.1
		500	None	24.1	25.4	35.1	45.9	40.0	47.0	55.0
			Block	29.6	28.2	36.7	42.9	39.5	46.5	52.1
			MD.W	12.8	21.7	24.1	22.1	26.3	29.2	37.4
			MD.WX.Full	23.7	24.8	35.3	44.9	35.7	41.4	46.4
			MD.X	9.9	10.5	15.2	23.5	22.3	27.3	31.9
			MD.X.Full	15.9	16.7	23.9	32.5	30.9	37.4	43.5
	ARCH.t	250	None	23.5	36.0	24.7	23.1	25.7	25.6	31.9
			Block	24.7	28.8	26.4	23.5	29.0	29.1	29.0
			MD.W	28.9	24.9	22.1	19.6	24.9	24.6	27.5
			MD.WX.Full	25.6	31.6	28.0	24.4	30.7	30.2	33.2
			MD.X	24.9	32.2	32.0	30.3	32.1	31.5	31.6
			MD.X.Full	51.6	55.5	53.5	51.3	51.5	50.9	51.5
		500	None	30.2	40.1	35.9	30.9	34.7	34.7	35.8
			Block	36.5	39.8	36.4	29.6	38.1	37.1	36.9
			MD.W	28.2	36.3	32.8	26.8	36.0	35.5	39.8
			MD.WX.Full	36.6	47.6	42.8	36.7	46.1	46.0	50.1
			MD.X	44.6	58.4	57.8	57.1	58.6	58.2	58.0
			MD.X.Full	81.2	83.1	82.1	80.9	80.5	80.5	80.6

Table 21: Rejection rates for various bad realized PIT estimation methods and various tests in the dynamic backtesting experiment. The data-generating mechanism for the losses is the GARCH.t process. Models are refitted after 10 days. Results are based on 1000 replications, with $\alpha_1 = 0.975$ *and* $\alpha_2 = 0.9995$.

n_2	$\widehat{F}$	n	ITT | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	PNS
250	EWMA.HS	250	None	8.7	9.6	13.7	21.1	14.3	21.0	30.3
			Block	15.2	17.9	18.8	20.9	21.7	26.2	33.6
			MD.W	23.4	17.3	17.5	16.3	17.2	17.9	22.9
			MD.WX.Full	10.7	14.8	17.8	23.5	16.3	19.2	26.2
			MD.X	5.8	5.8	6.7	7.8	6.8	8.0	12.3
			MD.X.Full	9.9	9.9	11.3	14.8	12.9	15.1	23.8
		500	None	7.2	7.9	15.5	31.1	15.2	26.7	44.7
			Block	26.4	19.7	24.4	29.9	24.3	33.1	46.7
			MD.W	9.8	21.3	22.0	20.0	22.2	24.0	34.1
			MD.WX.Full	12.6	18.6	24.4	32.8	21.7	26.5	38.8
			MD.X	7.5	10.1	11.1	13.9	11.4	14.8	24.1
			MD.X.Full	12.5	13.7	16.1	23.3	15.8	24.6	39.5
	HS	250	None	36.8	46.3	43.1	46.6	41.2	45.2	54.9
			Block	43.9	49.3	48.6	47.8	49.5	51.8	54.8
			MD.W	48.4	47.4	44.7	43.8	47.4	48.4	51.7
			MD.WX.Full	43.9	49.9	49.4	50.7	50.2	51.4	57.8
			MD.X	39.5	49.9	50.8	52.5	50.5	51.8	54.2
			MD.X.Full	71.4	74.4	75.0	75.1	74.7	75.4	77.9
		500	None	45.1	52.4	57.0	64.2	57.2	62.8	71.9
			Block	67.0	66.9	67.6	68.1	68.9	72.2	76.9
			MD.W	55.9	68.4	67.0	63.5	69.5	70.5	75.5
			MD.WX.Full	68.2	77.5	76.4	76.9	77.8	80.0	83.9
			MD.X	66.6	79.5	80.6	81.2	79.5	81.6	83.7
			MD.X.Full	94.2	94.8	94.9	95.3	94.8	95.1	96.0
500	EWMA.HS	250	None	7.7	7.3	7.5	10.4	8.3	10.7	16.2
			Block	11.3	12.9	11.9	10.2	14.4	15.9	19.8
			MD.W	18.0	12.0	10.6	9.6	11.4	11.7	13.1
			MD.WX.Full	8.7	9.5	10.8	13.5	9.9	11.4	14.6
			MD.X	5.4	4.8	5.2	5.6	5.6	6.0	7.3
			MD.X.Full	6.8	7.4	8.8	10.4	8.9	9.4	13.3
		500	None	4.6	4.5	6.7	10.0	5.9	9.8	17.2
			Block	19.9	11.6	13.6	14.6	13.0	15.3	19.6
			MD.W	5.9	11.0	12.9	11.7	11.1	11.5	15.7
			MD.WX.Full	9.1	11.1	12.9	16.2	11.2	12.5	16.7
			MD.X	7.6	7.8	8.2	9.3	7.8	8.7	11.2
			MD.X.Full	11.8	11.7	12.6	13.7	11.1	12.9	17.5
	HS	250	None	29.1	44.4	31.8	33.0	32.0	32.9	45.2
			Block	34.8	37.7	36.8	35.0	38.2	39.4	41.2
			MD.W	39.1	36.1	34.3	32.4	35.9	36.4	39.9
			MD.WX.Full	35.5	40.5	38.9	37.5	40.0	40.2	44.7
			MD.X	35.9	42.0	41.9	42.0	42.2	41.8	43.6
			MD.X.Full	66.9	69.7	68.6	67.7	68.1	68.0	69.4
		500	None	40.4	49.9	49.7	49.6	48.8	51.4	57.4
			Block	54.9	60.4	56.3	50.9	59.4	59.9	63.6
			MD.W	46.6	55.5	53.7	50.5	55.2	54.9	58.9
			MD.WX.Full	56.9	66.9	62.2	59.1	66.4	66.4	73.0
			MD.X	61.9	73.0	73.1	72.6	71.8	71.5	73.0
			MD.X.Full	93.5	94.1	94.2	94.4	93.9	94.5	94.7

Table 22: Rejection rates for various industrial realized PIT estimation methods and various tests in the dynamic backtesting experiment. The data-generating mechanism for the losses is the GARCH.t process. Models are refitted after 10 days. Results are based on 1000 replications, with $\alpha_1 = 0.975$ *and* $\alpha_2 = 0.9995$.

6.5 Experiment three: Asymmetric GARCH process

The experimental design is similar to Section 6.3.1, except that we now use GJR-GARCH of Glosten et al. (1993) as the data generating process. The motivation for this experiment is due to the fact that we observed low rejection rate for the EWMA.HS model in Section 6.4. We speculate that this is because the EWMA process is somewhat similar to the standard GARCH process.

We will now assume that the model for financial returns takes the form

$$R_t = \sigma_t Z_t, \quad \sigma_t^2 = \varphi_0 + (\varphi_1 + \omega_1 I_{\{R_{t-1}<0\}})R_{t-1}^2 + \theta_1 \sigma_{t-1}^2 \,, \tag{6.6}$$

where (Z_t) is an iid innovation distribution with zero mean and unit variance. We will assume that (Z_t) follows a (standardized) skewed Student-t distribution with ν degrees of freedom and skewness parameter γ. The simulated losses (L_t) are given by $L_t = -R_t$. We fit the model to the S&P index log-returns for the period January 2007 to December 2012 to obtain the parameter values $\varphi_1 = 6.398e-08$, $\omega_1 = 0.186$, $\theta_1 = 0.897$, $\nu = 8.043$ and $\gamma = 0.869$.

Note that the value of γ means that the distribution of (Z_t) is skewed to the left, and hence the distribution of L_t is skewed to the right. Also, the value of ω_1 means that volatility reacts more to negative market shocks. To ensure that the process is stationary with variance one, we set $\varphi_0 = 1 - \kappa$, where κ is the persistence of the process, and is given by $\kappa = \varphi_1 + \gamma_1 \mathrm{E}\left(Z^2 I_{\{Z \leqslant 0\}}\right) + \theta_1 \approx 0.998$.

The test results are given in Table 23. As we would have expect, the rejection rates of the tests for the industrial models are now larger when compared to those in Table 22, especially for the EWMA.HS model when $n = 500$. Note that the increase in rejection rates are due to the failure of the EWMA.HS and HS model in capturing the dynamic of the losses. Recall from Section 4.4.2 that the rejection rates for the HS model are not affected by the fatness in the tail of the data generating process.

n_2	$\widehat{F}$	n	ITT | test	B99	SP.U	SP.L	SP.E.200	SP.UL	SP.UE.200	PNS
250	EWMA.HS	250	None	12.8	12.4	17.9	26.4	16.5	24.1	35.7
			Block	23.8	27.4	26.5	27.8	29.0	33.2	40.1
			MD.W	35.3	29.2	28.6	26.6	28.4	29.4	33.2
			MD.WX.Full	14.7	19.3	22.8	29.2	20.6	23.2	31.8
			MD.X	9.7	10.7	12.3	14.9	12.7	15.0	19.0
			MD.X.Full	20.0	20.8	23.6	27.2	23.1	27.0	33.3
		500	None	8.4	11.6	19.5	33.0	17.1	28.0	47.8
			Block	35.5	31.4	34.6	36.1	33.6	40.5	53.8
			MD.W	18.6	34.4	34.6	30.3	34.9	36.5	48.2
			MD.WX.Full	18.0	23.8	29.4	38.7	26.0	30.3	41.8
			MD.X	12.9	16.0	18.4	21.5	17.9	21.6	30.8
			MD.X.Full	33.0	34.0	37.2	44.5	37.3	43.3	53.4
	HS	250	None	36.5	44.7	41.7	44.9	41.7	45.3	54.9
			Block	43.7	48.8	48.5	47.0	49.8	52.3	55.7
			MD.W	50.4	49.7	47.1	46.7	49.4	49.8	53.5
			MD.WX.Full	43.8	49.9	50.2	50.8	50.6	51.7	56.9
			MD.X	44.0	53.1	53.5	55.2	53.6	54.8	56.7
			MD.X.Full	79.4	80.5	81.4	81.4	79.8	80.5	83.2
		500	None	45.9	50.8	55.7	63.0	57.8	64.6	72.2
			Block	69.7	68.2	69.3	70.0	71.4	73.8	78.1
			MD.W	62.0	72.1	70.8	67.5	72.4	73.1	78.8
			MD.WX.Full	68.0	76.1	76.1	76.7	76.5	78.7	82.9
			MD.X	72.4	81.9	82.7	83.3	82.7	83.4	84.3
			MD.X.Full	96.3	97.0	96.9	97.2	96.8	97.2	97.6
500	EWMA.HS	250	None	8.0	8.3	9.5	13.1	10.8	13.4	17.9
			Block	16.6	19.3	17.1	15.6	20.5	21.9	27.0
			MD.W	26.2	21.7	19.0	17.1	21.3	21.4	24.4
			MD.WX.Full	11.7	13.7	15.0	16.3	13.0	13.9	17.5
			MD.X	8.1	10.2	10.1	11.1	9.7	11.0	12.5
			MD.X.Full	20.5	19.6	20.6	21.6	20.1	22.2	23.8
		500	None	4.9	5.6	7.4	12.1	8.1	10.7	15.8
			Block	30.9	22.9	23.8	21.4	23.9	26.0	30.4
			MD.W	17.4	28.0	27.2	22.7	27.6	28.5	33.7
			MD.WX.Full	13.4	16.0	17.4	19.5	16.5	18.2	21.9
			MD.X	11.4	15.8	16.4	17.0	15.4	16.2	18.7
			MD.X.Full	33.4	34.3	34.9	37.0	33.0	35.6	37.5
	HS	250	None	32.8	53.2	34.7	35.9	35.3	36.7	52.5
			Block	39.0	43.5	41.8	38.9	44.5	45.2	44.8
			MD.W	43.3	43.0	40.5	38.6	42.7	42.9	46.1
			MD.WX.Full	38.9	43.6	42.6	41.3	43.1	43.3	47.1
			MD.X	41.7	50.5	50.2	50.1	50.0	50.3	50.5
			MD.X.Full	79.2	82.3	80.1	79.4	80.2	79.9	80.9
		500	None	45.8	53.2	53.8	54.9	52.7	55.8	59.9
			Block	61.7	69.5	64.3	55.6	67.2	67.2	71.6
			MD.W	54.5	64.2	60.8	56.6	64.2	64.1	66.1
			MD.WX.Full	59.0	70.5	64.4	61.7	69.1	68.3	76.3
			MD.X	69.9	80.6	80.3	80.0	79.3	79.5	80.2
			MD.X.Full	97.2	97.8	97.5	97.5	97.3	97.5	97.4

Table 23: Rejection rates for various industrial realized PIT estimation methods and various tests in the dynamic backtesting experiment. The data-generating mechanism for the losses is the GJR-GARCH process with skewed Student innovations. Models are refitted after 10 days. Results are based on 1000 replications, with $\alpha_1 = 0.975$ *and* $\alpha_2 = 0.9995$.

Chapter 7 Elicitability theory and model selection

We end the thesis with a chapter on elicitability theory and model selection, focusing on the weighted scoring rule which we will describe later, since it bears close resemblance to the idea of weighted realized PIT values. In particular, we will compare and rank competing forecast distributions based on elicitability theory.

To be able to rank forecast distributions, we will need to assign scores to them, based on some function. In this chapter, we will use the convention that a smaller score is associated to a better model. For a realized loss variable $L_t \in \mathbb{R}$ with distribution F_t (which we assume to be continuous), suppose a forecaster quotes the predictive distribution $\widehat{F}_t$, then the expected score of the predictive distribution is $\mathrm{E}\left(R(\widehat{F}_t, L_t)\right)$, where $R(F, \cdot)$ is the scoring rule that takes value on the real line $\mathbb{R}$. The scoring rule $R(F, \cdot)$ is said to be proper if $\mathrm{E}\left(R(F_t, L_t)\right) \leqslant \mathrm{E}\left(R(\widehat{F}_t, L_t)\right)$, and strictly strictly if it is proper, and equality of the expectation implies that $\widehat{F}_t = F_t$. See for example, Gneiting & Raftery (2007) and the references therein for more details on different types of proper scoring rules.

In many practical situations, we may be more interested in the evaluation of single-valued point forecasts, rather than the entire forecast distribution. This is especially true in the regulatory setting, where the regulators do not have full information on the forecast distribution, but only have access to a limited set of data submitted by banks. See Gneiting (2011) for more details on evaluation of different types of point forecasts. In this chapter, we will focus on the evaluation of VaR.

7.1 Evaluation of VaR at a single level α

The Value at Risk at level α, denoted by $\mathrm{VaR}_{\alpha,t} = F_t^{-1}(\alpha)$, is known to be elicitable. This means that there exist some scoring function $S_\alpha(q, l)$ which takes input $q \in \mathbb{R}$ and $l \in \mathbb{R}$, such that $\mathrm{E}\left(S_\alpha(\mathrm{VaR}_{\alpha,t}, L_t)\right) \leqslant \mathrm{E}\left(S_\alpha(q, L_t)\right)$ for all q. When the inequality holds, we say that the function $S_\alpha(q, l)$ is consistent for $\mathrm{VaR}_{\alpha,t}$, and it is strictly consistent if it is consistent, and equality of the expectation implies that $q = \mathrm{VaR}_{\alpha,t}$. In other words, if $S_\alpha(q, l)$ is strictly consistent, we can obtain $\mathrm{VaR}_{\alpha,t}$

by minimizing the expected score with respect to q, with

$$\mathrm{VaR}_{\alpha,t} = \arg\min_{q\in\mathbb{R}} \mathrm{E}\left(S_\alpha(q, L_t)\right). \tag{7.1}$$

Up to mild regularity conditions, the scoring function $S_\alpha(q, l)$ is consistent for VaR at level α if and only if it is of the form

$$S_\alpha(q, l) = I_{\{l<q\}}(1-\alpha)\big(s(q) - s(l)\big) + I_{\{l\geqslant q\}}\alpha\big(s(l) - s(q)\big), \tag{7.2}$$

for some non-decreasing function s. It is strictly consistent if s is strictly increasing. See Thomson (1979) and Saerens (2000). Note that if $S_\alpha(q, l)$ is strictly consistent, then linear transformations of $S_\alpha(q, l)$, denoted by

$$\tilde{S}_\alpha(q, l) = k\, S_\alpha(q, l) + a(l), \tag{7.3}$$

for some constant $k > 0$ and some function a, are also strictly consistent. See Fissler & Ziegel (2015).

The scoring function $S_\alpha(q, l)$ in (7.2) depends on the choice of function s, which is not uniquely determine, since the only requirement is that we require s to be non-decreasing. A popular choice would be to set $s(x) = x$, which leads to

$$S_\alpha(q, l) = I_{\{l<q\}}(1-\alpha)\big(q - l\big) + I_{\{l\geqslant q\}}\alpha\big(l - q\big). \tag{7.4}$$

In this case $S_\alpha(q, l)$ is known as the asymmetric piecewise linear (APL) scoring function. Ehm et al. (2016) have shown that (7.2) can be rewritten in the Choquet-type mixture representations

$$S_\alpha(q, l) = \int_{-\infty}^{\infty} S_{\alpha,\beta}(q, l)\,\mathrm{d}M(\beta), \tag{7.5}$$

where M is some non-negative measure and

$$S_{\alpha,\beta}(q, l) = I_{\{l\leqslant\beta<q\}}(1-\alpha) + I_{\{q\leqslant\beta<l\}}\alpha. \tag{7.6}$$

$S_{\alpha,\beta}(q, l)$ is known as the elementary quantile scoring function at point β, and $S_\alpha(q, l)$ is obtained by taking a weighted integral of the elementary quantile scoring function. In particular, when M is the Lebesgue measure, $S_\alpha(q, l)$ in (7.5) becomes the APL scoring function. Hence, (7.5) changes the problem of choosing the function s to one of choosing the weight measure M.

In this section, we will only consider the APL scoring function. It is easy to show that the APL scoring function in (7.4) satisfies

$$\left.\frac{\partial \mathrm{E}\left(S_\alpha(q, L_t)\right)}{\partial q}\right|_{q=F^{-1}(\alpha)} = 0\,. \tag{7.7}$$

Hence, according to Lambert et al. (2008) Definition 4 and Definition 5, the APL scoring function is accuracy-rewarding. This means that for some $q^* \in \mathbb{R}$,

$$\mathrm{E}\left(S_\alpha(q^*, L_t)\right) < \mathrm{E}\left(S_\alpha(q, L_t)\right) \tag{7.8}$$

when either $q < q^* < F_t^{-1}(\alpha)$ or $F_t^{-1}(\alpha) < q^* < q$. Note that if $S_\alpha(q, L_t)$ is accuracy-rewarding, it is easy to show that the linear transformation $\tilde{S}_\alpha(q, L_t)$ in (7.3) is accuracy-rewarding as well.

For a series of realized loss variables $(L_1, \ldots, L_n)$ with corresponding m competing VaR forecast estimates $(\widehat{\mathrm{VaR}}_{\alpha,1,k}, \ldots, \widehat{\mathrm{VaR}}_{\alpha,n,k})$ for $k = 1, \ldots, m$, the accuracy-rewarding property of APL implies that we can evaluate the VaR forecast performance by computing the average score $\bar{S}_{\alpha,k} = \frac{1}{n}\sum_{t=1}^{n} S_\alpha(\widehat{\mathrm{VaR}}_{\alpha,t,k}, L_t)$, and the forecast model which have the lowest average score should be preferred. If we wish to find the best model for a given confidence level, we could use, for example, the Diebold & Mariano (1995) test.

7.2 Diebold Mariano test

In this section we will summarize the Diebold & Mariano (1995) test with modifications suggested by Harvey et al. (1997), since the test is used extensively in the simulation studies later.

For $t = 1, \ldots, n$, suppose we have a series of score (S_t) computed using some forecast distribution $(\widehat{F}_t)$ for (L_t), and another series of benchmark score $(S_{0,t})$ computed using some benchmark distribution $(\widehat{F}_{0,t})$ for (L_t). We denote the score difference by $D_t = S_t - S_{0,t}$ and its sample mean by $\bar{D} = \frac{1}{n}\sum_{t=1}^{n} D_t$. We then construct the hypothesis

$$H_0: \quad \mathrm{E}\left(\bar{D}\right) \leqslant 0 \;\; \text{vs.} \;\; H_1 : \mathrm{E}\left(\bar{D}\right) > 0\,. \tag{7.9}$$

Under the assumption that the autocovariance function $\gamma(h) = \mathrm{cov}(D_{t+h}, D_t)$ equals

zero for $h \geqslant c$ for some cutoff point c, the variance of $\bar{D}$ is given by

$$\operatorname{var}(\bar{D}) = \frac{1}{n}\left(\gamma(0) + \frac{2}{n}\sum_{h=1}^{c-1}(n-h)\,\gamma(h)\right). \tag{7.10}$$

To test the hypothesis in (7.9), the Diebold Mariano (DM) test is based on the Z-test statistic

$$Z_{\mathrm{DB}} = \frac{\bar{D}}{\hat{\sigma}_D}, \tag{7.11}$$

which is compared to a standard normal distribution, where $\hat{\sigma}_D^2$ is some estimator for $\operatorname{var}(\bar{D})$ in (7.10). Diebold and Mariano propose using the estimator

$$\hat{\sigma}_D^2 = \frac{1}{n}\left(\hat{\gamma}^*(0) + \frac{2}{n}\sum_{h=1}^{c-1}(n-h)\,\hat{\gamma}^*(h)\right), \tag{7.12}$$

with $\hat{\gamma}^*(h) = \frac{n}{n-h}\hat{\gamma}(h)$, where $\hat{\gamma}(h)$ is the sample ACF in (5.2). The DM test is known to have poor size performance for moderate sample size n, especially when c is large. To improve the finite-sample size performance of DM test, Harvey et al. (1997) proposes to use the test statistic

$$T_{\mathrm{HLN}} = \left(\frac{n+1-2c+n^{-1}c(c-1)}{n}\right)^{1/2} Z_{\mathrm{DB}}, \tag{7.13}$$

which is compared to a Student-t distribution with $(n-1)$ degrees of freedom, where Z_{DB} is the DM test statistic given in (7.11). In the following simulation studies, we will base our tests using the test statistic T_{HLN} in (7.13). To obtain a test of approximately size κ, we reject H_0 in (7.9) when $T_{\mathrm{HLN}} > F_{n-1}(1-\kappa)$, where F_ν denotes the distribution of an ordinary Student-t distribution with ν degrees of freedom.

7.3 Evaluation of the weighted integral of VaR at different levels using a weighted scoring rule

Suppose we wish to evaluate the vector $\widehat{\mathrm{VaR}}_{\boldsymbol{\alpha},t} = (\widehat{\mathrm{VaR}}_{\alpha_1,t}, \ldots, \widehat{\mathrm{VaR}}_{\alpha_N,t})$ for $t = 1, \ldots, n$. Following the definition in Lambert et al. (2008), the score function for a vector of VaR estimates of size N, denoted as $S_{\boldsymbol{\alpha}}(q_1, \ldots, q_N, l)$, taking inputs $q_i \in \mathbb{R}$ for $i = 1, \ldots, N$ and $l \in \mathbb{R}$, is accuracy-rewarding if $\mathrm{E}\left(S_{\boldsymbol{\alpha}}(q_1^*, \ldots, q_N^*, L_t)\right) < \mathrm{E}\left(S_{\boldsymbol{\alpha}}(q_1, \ldots, q_N, L_t)\right)$ when for all i, either $q_i < q_i^* < F_t^{-1}(\alpha_i)$ or $F_t^{-1}(\alpha_i) < q_i^* < q_i$. According to Theorem 5 in Lambert et al. (2008), the sum of accuracy-rewarding

score functions is accuracy-rewarding, i.e.

$$S_{\boldsymbol{\alpha}}(q_1, \ldots, q_N, l) = \sum_{i=1}^{N} k_i \, S_{\alpha_i}(q_i, l) \,, \quad k_i > 0 \,, \tag{7.14}$$

is accuracy-rewarding if $S_{\alpha_i}(q_i, l)$, for $i = 1, \ldots, N$, are accuracy-rewarding. We will consider the APL scoring function, where $S_{\alpha_i}(q_i, l)$ is defined in (7.4).

To evaluate the forecast distribution $\widehat{F}_t$ for loss L_t in the range $[\alpha_1, \alpha_N]$, we set $q_i = \widehat{F}_t^{\leftarrow}(\alpha_i)$, for $i = 1, \ldots, N$, to be the generalized inverse of the forecast distribution $\widehat{F}_t$, where the generalized inverse function is defined in (3.2), and consider the sequence of evenly spaced levels $\alpha_1, \ldots, \alpha_N$. In the continuous limit as $N \to \infty$, we obtain the weighted scoring rule

$$R_g(\widehat{F}_t, L_t) = \int_{\alpha_1}^{\alpha_N} g(u) \, S_u\big(\widehat{F}_t^{\leftarrow}(u), L_t\big) \, \mathrm{d}u \,, \tag{7.15}$$

where g is some weight function satisfying Assumption 3.1. Note that when $S_\alpha(q, l)$ is the APL scoring function, the weighted scoring rule (7.15) is similar to the weighted version of the continuous ranked probability score (CRPS) defined in Gneiting & Ranjan (2011). The CRPS is a proper scoring rule, and is defined as

$$\begin{aligned} \mathrm{CRPS}(\widehat{F}_t, L_t) &= \int_{-\infty}^{\infty} \big(\widehat{F}_t(x) - I_{\{L_t \leqslant x\}}\big)^2 \, \mathrm{d}x \\ &= \int_0^1 \mathrm{QS}_u(\widehat{F}_t^{-1}(u), L_t) \, \mathrm{d}u \,, \end{aligned} \tag{7.16}$$

and the weighted version is $\int_0^1 g(u) \, \mathrm{QS}_u(\widehat{F}_t^{-1}(u), L_t) \, \mathrm{d}u$, for some non-negative weight function g, and $\mathrm{QS}_u(\widehat{F}_t^{-1}(u), L_t) = 2(I_{\{L_t \leqslant \widehat{F}_t^{-1}(u)\}} - u)(\widehat{F}_t^{-1}(u) - L_t)$. The main difference between $R_g(\widehat{F}_t, L_t)$ in (7.15) and the weighted CRPS is in the range which we take the integral.

7.4 Evaluation of VaR using weighted scoring rule in the case of limited data

In order to evaluate the weighted scoring rule (7.15), we require the knowledge of the forecast distribution $\widehat{F}_t$, which the tester may not have access to. This is the case for example when the tester is the regulator. Clearly, elicitability theory is much less useful in such cases. Nevertheless, the regulators may have access to the realized losses (L_t), and may have some views on suitable models for (L_t). In this

section, we assume that the tester have access to a set of realized PIT values, with the corresponding realized losses (L_t), and two sets of VaR estimates at the 97.5% and 99% level. The tester will try to create a proxy model for the (unknown) forecast model $\widehat{F}_t$ used by the banks using these sets of data, and compare the proxy model with a benchmark model.

The proxy model should be fairly flexible to allow for a wide range of possible forecast models $\widehat{F}_t$. For simplicity, we will construct a proxy model by assuming that (L_t) has the structure

$$L_t = \sigma_t Z_t \,, \tag{7.17}$$

where σ_t is a pre-visible constant, and (Z_t) are iid innovations with mean zero and unit variance. We denote by F_t^* the proxy distribution for $\widehat{F}_t$, and we will assume that (Z_t) are iid skewed Student-t with ν degrees of freedom and shape parameter ξ, scaled to have zero mean and unit variance. We denote the distribution function of the standardized skewed Student-t by $\tilde{F}_{\nu,\xi}$, and we set

$$\sigma_t = \frac{L_t}{\tilde{F}_{\nu,\xi}^{-1}(P_t)} \,, \tag{7.18}$$

and the proxy model is given by

$$F_t^*(x) = \tilde{F}_{\nu,\xi}\left(\frac{x}{\sigma_t}\right) . \tag{7.19}$$

We can then solve for ν and ξ simultaneously using the equations

$$\frac{\widehat{\mathrm{VaR}}_{0.975,t}}{L_t} - \frac{\tilde{F}_{\nu,\xi}^{-1}(0.975)}{\tilde{F}_{\nu,\xi}^{-1}(P_t)} = 0 \,, \tag{7.20}$$

$$\frac{\widehat{\mathrm{VaR}}_{0.99,t}}{L_t} - \frac{\tilde{F}_{\nu,\xi}^{-1}(0.99)}{\tilde{F}_{\nu,\xi}^{-1}(P_t)} = 0 \,. \tag{7.21}$$

We will refer to the weighted scoring rule in (7.15) where we replace $\widehat{F}_t$ with the proxy model F_t^* as the proxy weighted scoring rule. Note that we could consider other form of proxy models. For example, we could assume (Z_t) to have a generalized hyperbolic distribution, we could use a semi-parametric model based on Extreme Value Theory.

7.4.1 Simulation studies

We will now conduct simulations studies to better understand the scoring function and weighted scoring rule in (7.2) and (7.15), and the proxy weighted scoring rule described in Section 7.4. We will conduct three separate experiments.

In the first experiment, we will confirm the accuracy-rewarding property of the APL score at a single level α. The structure of the experiment will be based on a static data generating process, where the forecast models will not contain parameter estimation error.

The structure of the second experiment will also be based on a static data generating process, however, the forecast models will now contain parameter estimation error. The structure of the third experiment will be similar to the second experiment, except that the data generating process will be based on a GARCH process. For each of the second and third experiment, we will analyze the performance of the APL score in (7.2) at a single level, the weighted scoring rule in (7.15) based on the APL scoring function, and the proxy weighted scoring rule described in Section 7.4 based on the APL scoring function.

7.4.2 Experiment one: Static DGP, no parameter estimation error

The motivation of the first experiment is to confirm the accuracy-rewarding property of the APL score at a single level α.

The structure of the experiment is similar to Section 4.1.1, except that the true distribution F_t is always the skewed Student-t distribution of Fernandez & Steel (1998) with 3 degrees of freedom and a skewness parameter $\gamma = 1.2$, standardized to have zero mean and unit variance.

For the forecast models $\widehat{F}_t$, we consider the Student-t distribution with three and five degrees of freedom, scaled to have unit variance, denoted by $t3$ and $t5$, and the standard normal distribution.

For each simulation, we compute the average score $\bar{S}_\alpha = \frac{1}{n}\sum_{t=1}^{n} S_\alpha(\widehat{\text{VaR}}_{\alpha,t}, L_t)$. We also compute the average benchmark score using the VaR of the true model, which we denote by $\bar{S}_{\alpha,0} = \frac{1}{n}\sum_{t=1}^{n} S_\alpha(\text{VaR}_{\alpha,t}, L_t)$. The score difference and its sample mean are denoted respectively by $D_{\alpha,t} = S_\alpha(\widehat{\text{VaR}}_{\alpha,t}, L_t) - S_\alpha(\text{VaR}_{\alpha,t}, L_t)$ and $\bar{D}_\alpha = \frac{1}{n}\sum_{t=1}^{n} D_{\alpha,t}$. We repeat $1,000$ times to obtain an estimate of the mean and standard deviation of $\bar{D}_\alpha$, which is shown in Table 24. In addition, we also show the rejection rate when the Diebold & Mariano (1995) test with modifications

suggested by Harvey et al. (1997), as described in Section 7.2, is applied to the series $(D_{\alpha,t})$. The null and alternate hypothesis of the Diebold Mariano (DM) test is

$$H_0: \quad \mathrm{E}\,(\bar{D}_\alpha) \leqslant 0 \;\; \text{vs.} \;\; H_1 : \mathrm{E}\,(\bar{D}_\alpha) > 0\,. \tag{7.22}$$

Note that a one sample t-test will give similar results to the DM test for this experiment, since the data generating process and forecast distributions are serially independent.

From Table 24, we see that for the levels $\alpha = 0.975$ and $\alpha = 0.99$, the ranking of the mean of $\bar{D}_\alpha$ is consistent with the results in Table 3, in the sense that VaR forecasts that are further away from the true VaR will give a larger score, which confirms the accuracy-rewarding property of the APL scoring function. For example, at level $\alpha = 0.975$, the VaR of the standardized Student-$t5$ distribution is closest to the VaR of the standardized skewed Student-$t5$ distribution, whereas the VaR of the standardized Student-$t3$ is furthest away. Also, the rankings of the standard deviation of $\bar{D}_\alpha$ and the DM rejection rate is the same as the rankings of the mean of $\bar{D}_\alpha$.

Result Type	$\widehat{F}_t \mid \alpha$	0.9750	0.9900	0.9995
Mean of $\bar{D}_\alpha$	$t3$	0.07 (3)	0.07 (1)	0.02 (1)
	$t5$	0.00 (1)	0.08 (2)	0.19 (2)
	Normal	0.01 (2)	0.25 (3)	0.93 (3)
Standard Deviation of $\bar{D}_\alpha$	$t3$	3.34 (3)	3.88 (1)	2.34 (1)
	$t5$	0.83 (1)	4.05 (2)	8.15 (2)
	Normal	1.33 (2)	7.30 (3)	18.09 (3)
DM Rejection rate	$t3$	12.3 (3)	10.1 (1)	0.6 (1)
	$t5$	5.0 (1)	10.7 (2)	2.8 (2)
	Normal	6.1 (2)	24.2 (3)	41.4 (3)

Table 24: The estimated mean and standard deviation of $\bar{D}_\alpha$, and the Diebold Mariano test rejection rate, at varying levels α. The DGP is based on a skewed Student-t distribution. Results are in % and obtained using 1000 simulations. The values in the brackets are the rankings in ascending order.

7.4.3 Experiment two: Static DGP, with parameter estimation error

The second experiment is similar to the previous experiment, except that we now include parameter estimation error. The true distribution F_t is the skewed Student-t distribution with 3 degrees of freedom and a skewness parameter $\gamma = 1.2$, standardized to have zero mean and unit variance.

The forecast models $\widehat{F}_t$ will assume that (L_t) has the structural form

$$L_t = \mu + \sigma Z_t \,, \tag{7.23}$$

where Z_t are iid innovations with zero mean and unit variance. We consider three possibilities for the distribution of Z_t for the forecast models, the normal distribution, Student-t distribution (denoted as t), skewed Student-t distribution (denoted as st), all standardized to have zero mean and unit variance. Parameter estimates are obtained using a rolling window of size $n_2 = 500$. Additionally, we also consider the historical simulation (HS) model in (4.21), with rolling window of size $n_2 = 250$ and $n_2 = 500$, which we denote by HS.250 and HS.500 respectively. For the HS models, we estimate $\widehat{\mathrm{VaR}}_{\alpha,t}$ using (4.25).

We first look at the results of the APL score at a single level, which are shown in Table 25. We observe that the $\bar{D}_\alpha$ of the skewed Student-t model, which assumes the correct structural form, always has the lowest mean and standard deviation. However, this ranking is not reflected in the DM rejection rate at level $\alpha = 0.99$ and $\alpha = 0.9995$. In particular, the skewed Student-t model has a much higher rejection rate at level $\alpha = 0.9995$ compared to the Student-t and HS models. This is because, in the presence of parameter estimation error, at very high levels α, the variance of the series $(D_{\alpha,t})$ of certain forecast models (for example, the Student-t and HS models in this experiment) is very large, which results in a small DM test statistic. In other words, even though the accuracy-rewarding property still holds, when the level α is large, the DM test cannot rank the forecast models reliably.

We now look at the results of the weighted APL score (7.15). We will consider the uniform, linear and exponential weight function described in Section 3.4.3, normalized to have unit area. For each simulation, we estimate the average score $\bar{R}_g = \frac{1}{n}\sum_{t=1}^{n} R_g(\widehat{F}_t, L_t)$. We also estimate the average benchmark score using the

Result Type	$\widehat{F}_t \mid \alpha$	0.9750	0.9900	0.9995
Mean of $\bar{D}_\alpha$	*st*	0.05 (1)	0.06 (1)	0.04 (1)
	t	0.18 (5)	0.15 (3)	0.06 (2)
	Normal	0.17 (3)	0.41 (5)	1.06 (5)
	HS.500	0.09 (2)	0.13 (2)	0.22 (3)
	HS.250	0.17 (3)	0.22 (4)	0.45 (4)
Standard Deviation of $\bar{D}_\alpha$	*st*	2.73 (1)	3.21 (1)	2.63 (1)
	t	5.30 (5)	5.55 (3)	3.88 (2)
	Normal	4.70 (3)	9.14 (5)	19.34 (5)
	HS.500	3.52 (2)	4.48 (2)	8.39 (3)
	HS.250	5.02 (4)	6.19 (4)	12.48 (4)
DM Rejection Rate	*st*	14.4 (1)	15.6 (2)	23.9 (4)
	t	23.1 (4)	14.3 (1)	3.9 (1)
	Normal	21.9 (3)	34.6 (5)	52.7 (5)
	HS.500	19.6 (2)	21.6 (3)	6.7 (2)
	HS.250	27.9 (5)	28.3 (4)	9.2 (3)

Table 25: The estimated mean and standard deviation of $\bar{D}_\alpha$, and the Diebold Mariano test rejection rate, at varying levels α. The DGP is based on a skewed Student-t distribution. Results are in % and obtained using 1000 simulations. The values in the brackets are the rankings in ascending order.

true model, which we denote by $\bar{R}_{g,0} = \frac{1}{n}\sum_{t=1}^{n} R_g(F_t, L_t)$. The score difference and its sample mean is denoted respectively by $D_{g,t} = R_g(\widehat{F}_t, L_t) - R_g(F_t, L_t)$ and $\bar{D}_g = \frac{1}{n}\sum_{t=1}^{n} D_{g,t}$. We repeat $1,000$ times to obtain an estimate of the mean and standard deviation of $\bar{D}_g$, as well as the rejection rate of the DM test applied to the series $(D_{g,t})$, which is shown in Table 26.

From Table 26, we see that the accuracy-rewarding property continues to hold true, in the sense that the mean of the $\bar{D}_g$ for the skewed Student-t forecast model is always the smallest, and models that we expect to poorly estimate the tail of the distribution have relatively larger score. By taking a weighted integral of the APL score, the score difference $(D_{g,t})$ can take on a wider range of values, and its variance can be better estimated, which leads to a more reliable DB test. In particular, for the uniform weighting, the ranking of the DB test rejection rates is similar to the ranking of the mean of $\bar{D}_g$, with exception to the case when $\widehat{F}_t$ is the HS models (for example, the rejection rate when $\widehat{F}_t$ is HS.250 is larger than when $\widehat{F}_t$ is normal, even though HS.250 has a lower score). The reliability of the DB test gets worse as we place more weight in the tail.

Result Type	$\widehat{F}_t$ | weight	Uniform	Linear	Exponential
Mean of $\bar{D}_g$	st	0.05 (1)	0.05 (1)	0.05 (1)
	t	0.15 (3)	0.14 (2)	0.12 (2)
	Normal	0.41 (5)	0.54 (5)	0.70 (5)
	HS.500	0.12 (2)	0.14 (2)	0.17 (3)
	HS.250	0.24 (4)	0.28 (4)	0.33 (4)
Standard Deviation of $\bar{D}_g$	st	2.65 (1)	2.76 (1)	2.84 (1)
	t	4.76 (3)	4.72 (3)	4.64 (3)
	Normal	8.12 (5)	10.18 (5)	12.62 (5)
	HS.500	3.53 (2)	3.87 (2)	4.44 (2)
	HS.250	5.01 (4)	5.52 (4)	6.43 (4)
DM Rejection rate	st	18.2 (1)	16.8 (2)	16.5 (2)
	t	19.5 (2)	14.1 (1)	8.6 (1)
	Normal	45.6 (4)	51.3 (4)	55.9 (5)
	HS.500	35.6 (3)	37.9 (3)	36.2 (3)
	HS.250	53.0 (5)	53.8 (5)	47.2 (4)

Table 26: The estimated mean and standard deviation of $\bar{D}_g$, and the Diebold Mariano test rejection rate, at varying weight functions. The DGP is based on a skewed Student-t distribution. Results are in % and obtained using 1000 simulations. The values in the brackets are the rankings in ascending order.

We now consider the proxy weighted scoring rule based on Section 7.4. For each simulation, we estimate the proxy average score $\bar{R}^*_g = \frac{1}{n}\sum_{t=1}^{n} R_g(F^*_t, L_t)$, where F^*_t is the proxy model estimated using methods described in Section 7.4. The required set of data, i.e. the realized PIT values, and two sets of VaR estimates at the 97.5% and 99% level, is computed using the forecast distribution $\widehat{F}_t$, where $\widehat{F}_t$ is estimated using a rolling window procedure of size $n_2 = 500$. For the HS model we consider rolling window of size $n_2 = 250$ and $n_2 = 500$, which we denote by HS.250 and HS.500 respectively. We also compute the proxy average benchmark score using the true model, which we denote by $\bar{R}^*_{g,0} = \frac{1}{n}\sum_{t=1}^{n} R_g(F^*_{t,0}, L_t)$, where $F^*_{t,0}$ is the proxy model estimated using the true realized PIT values and true VaR. The proxy score difference and its sample mean are denoted respectively by $D^*_{g,t} = R_g(F^*_t, L_t) - R_g(F^*_{t,0}, L_t)$ and $\bar{D}^*_g = \frac{1}{n}\sum_{t=1}^{n} D^*_{g,t}$. We repeat $1,000$ times to obtain an estimate of the mean and standard deviation of $\bar{D}^*_g$, as well as the rejection rate of the DM test applied to the series $(D^*_{g,t})$, which is shown in Table 27.

From Table 27, we see that the results when $\widehat{F}_t$ is normal, Student-t, and skewed Student-t is very similar to those in Table 26. Compared to Table 26, the HS models

have much worse score and higher DM test rejection rate. This is because the proxy model structure that we use cannot capture the HS structure properly. Nevertheless, it is clear that with a flexible proxy distribution for L_t, and with enough data, we can still compare models using the weighted scoring rule (7.15), in the absence of the knowledge of the full forecast distribution $\widehat{F}_t$.

Result Type	$\widehat{F}_t$ | weight	Uniform	Linear	Exponential
Mean of $\bar{D}_g^*$	st	0.05 (1)	0.05 (1)	0.05 (1)
	t	0.16 (2)	0.14 (2)	0.12 (2)
	Normal	0.66 (5)	0.81 (5)	0.99 (5)
	HS.500	0.30 (3)	0.38 (3)	0.49 (3)
	HS.250	0.39 (4)	0.49 (4)	0.62 (4)
Standard Deviation of $\bar{D}_g^*$	st	3.38 (1)	3.62 (1)	3.93 (1)
	t	5.88 (2)	6.11 (2)	6.43 (2)
	Normal	16.42 (5)	19.30 (5)	22.88 (5)
	HS.500	6.09 (3)	7.50 (3)	9.87 (3)
	HS.250	6.79 (4)	8.16 (4)	10.25 (4)
DM Rejection rate	st	18.3 (1)	18.0 (2)	18.3 (2)
	t	19.7 (2)	13.2 (1)	7.5 (1)
	Normal	45.3 (3)	51.4 (3)	55.4 (3)
	HS.500	65.3 (4)	69.0 (4)	66.8 (4)
	HS.250	71.3 (5)	75.6 (5)	76.7 (5)

Table 27: The estimated mean and standard deviation of $\bar{D}_g^*$, and the Diebold Mariano test rejection rate, at varying weight functions, based on the weighted scoring rule when we replace the forecast model $\widehat{F}_t$ with the proxy model F_t^* estimated using available data sets. The DGP is based on a skewed Student-t distribution. Results are in % and obtained using 1000 simulations. The values in the brackets are the rankings in ascending order.

7.4.4 Experiment three: GARCH DGP, with parameter estimation error

The third experiment is similar to the experiments in the previous section, except that (L_t) is now simulated from a GARCH(1,1) process with Student-t innovations. The parameter estimates for the data generating process and the forecast models that we consider is the same as those in Section 6.3.1. We set the sample size $n = 1000$, and estimate parameters using rolling window of size $n_2 = 500$. For the HS and EWMA.HS model, we also consider $n_2 = 250$. We denote by HS.250 and EWMA.HS.250 the case when $n_2 = 250$, and similarly HS.500 and EWMA.HS.500

the case when $n_2 = 500$. The experiment is repeated 1000 times.

We first consider the case of the single level APL score. Similar to the previous section, we compute the APL average score difference $\bar{D}_\alpha$ using the forecast model and true model (the Oracle), and estimate the mean and standard deviation of $\bar{D}_\alpha$ via simulations. We also estimate the DM test rejection rates. The results are shown in Table 28. The key observation here is that models which misspecified the dynamics, namely the ARCH.t and HS models, are penalized much more heavily than the models that underestimates the tail of the innovation distribution, namely the GARCH.norm and EWMA.HS.250 models, resulting in a much higher DM test rejection rate.

Recall the results in Table 21, where the rejection rate of the binomial test at level $\alpha = 0.99$ for the ARCH.t model is only slightly higher than for the GARCH.norm model. To understand the cause of the difference in rejection rate of the DM test using APL score versus the binomial test, we refer to Figure 12, which plots the (L_t) of a particular simulation, and the corresponding VaR estimates of the Oracle, GARCH.norm and ARCH.t forecast models. Notice that most of the time the absolute difference between the VaR estimates of the ARCH.t and Oracle model is much larger than the absolute difference between the GARCH.norm and Oracle model. Regardless of direction, a large difference between the VaR estimates and true VaR will result in a large APL score, hence, the APL score is very effective in detecting misspecification in dynamics. On the other hand, the binomial test rejection rate depends on the average exception rate. The ARCH.t model has periods with high exception rate and periods with low exception rate. When taking averages, the cancellation effect results in a much smaller test rejection rate.

Next, we look at the results of the weighted APL score in (7.15). Similar to the previous section, we consider the uniform, linear and exponential weight function normalized to have unit area, and estimate the mean and standard deviation of the average weighted score difference $\bar{D}_g$, as well as the DM test rejection rates via simulations. The results in shown in Table 29.

We see that the accuracy-rewarding property still holds, in the sense that the mean of the $\bar{D}_g$ for the GARCH.t forecast model is always the smallest. Similar to the

Result Type	$\widehat{F}_t \mid \alpha$	0.9750	0.9900	0.9995
Mean of $\bar{D}_\alpha$	GARCH.t	0.0007 (1)	0.0005 (1)	0.0002 (1)
	GARCH.norm	0.0010 (2)	0.0014 (2)	0.0021 (3)
	ARCH.t	0.0214 (5)	0.0145 (5)	0.0023 (5)
	EWMA.HS.500	0.0020 (3)	0.0018 (3)	0.0010 (2)
	HS.500	0.0326 (7)	0.0207 (7)	0.0054 (6)
	EWMA.HS.250	0.0026 (4)	0.0025 (4)	0.0022 (4)
	HS.250	0.0265 (6)	0.0169 (6)	0.0078 (7)
Standard Deviation of $\bar{D}_\alpha$	GARCH.t	0.0331 (1)	0.0295 (1)	0.0147 (1)
	GARCH.norm	0.0387 (2)	0.0504 (3)	0.0620 (5)
	ARCH.t	0.2060 (5)	0.1638 (5)	0.0404 (3)
	EWMA.HS.500	0.0550 (3)	0.0497 (2)	0.0372 (2)
	HS.500	0.2456 (7)	0.1864 (7)	0.1088 (6)
	EWMA.HS.250	0.0634 (4)	0.0576 (4)	0.0592 (4)
	HS.250	0.2217 (6)	0.1758 (6)	0.1378 (7)
DM Rejection rate	GARCH.t	15.6 (2)	13.5 (2)	16.9 (4)
	GARCH.norm	14.4 (1)	11.8 (1)	9.1 (2)
	ARCH.t	86.8 (5)	76.3 (5)	61.9 (7)
	EWMA.HS.500	27.0 (3)	34.3 (3)	8.7 (1)
	HS.500	91.1 (7)	83.2 (7)	22.7 (5)
	EWMA.HS.250	34.3 (4)	38.4 (4)	11.6 (3)
	HS.250	89.8 (6)	80.0 (6)	30.9 (6)

Table 28: The estimated mean and standard deviation of $\bar{D}_\alpha$, and the Diebold Mariano test rejection rate, at varying levels α. The DGP is based on a GARCH process with Student-t innovations. Results are in % and obtained using 1000 simulations. The values in the brackets are the rankings in ascending order.

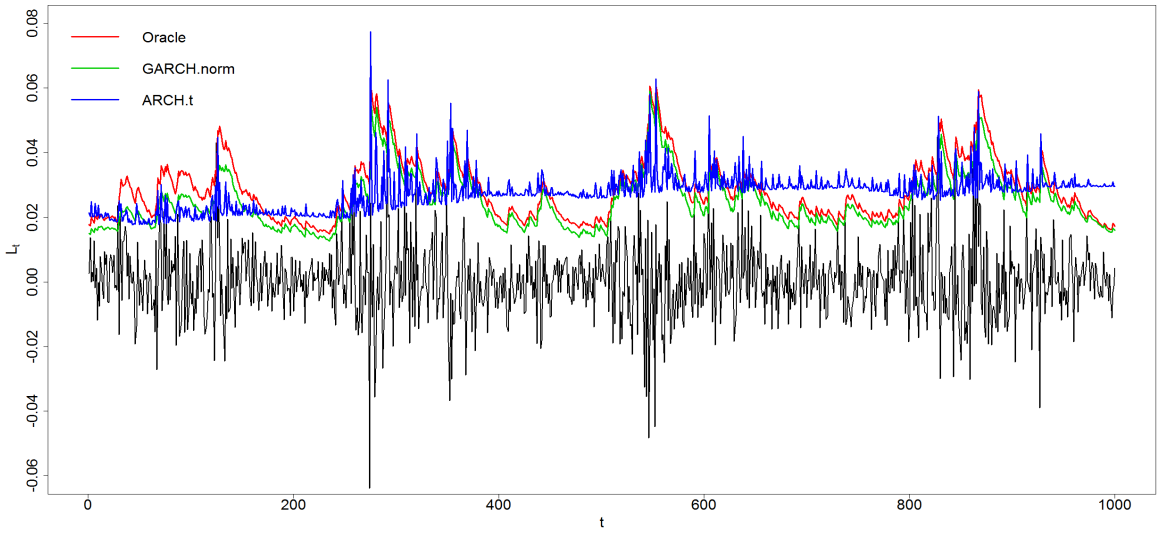

Figure 12: L_t of a particular realization (black line), and the corresponding VaR estimates of the Oracle, GARCH.norm and ARCH.t forecast models.

observations in Table 28 for the single level APL scores, models that misspecify the dynamics (ARCH.t and HS models) are penalized more heavily, resulting in a much larger DM test rejection rate. We also observe that the DM test for the weighted APL scores is more reliable compared to the DM test for APL scores at a single level, in the sense that GARCH.t forecast model always has the lowest rejection rate.

One important observation here is that the DB test rejection rates for the HS and EWMA.HS model are much higher than those observed in Table 22 in Section 6.4. Recall also the results in Table 12 from Section 4.4.2 that most of the static tests, with the exception of M.true, have great difficulty in rejecting the HS model when estimation window $n_2 = 500$. This suggest that bias-based tests as discussed in Section 3.2 are more effective in rejecting the HS model than exception-based test.

The DB test for APL score difference has similar characteristics to an absolute bias-based test since the APL score difference is large when the absolute difference between the estimated VaR and true VaR (i.e. the absolute bias) is large, and vice versa. Hence, we would expect that it would be able to detect models which systematically underestimates the true VaR. We cannot test the observed weighted APL scores of the forecast distribution directly, since the distribution of the scores depends on the true loss distribution F_t, which is not known. However, if our goal is to select a suitable forecast model from a pool of models, making decisions based on

the weighted APL score ranking will help us to avoid models with wrong dynamics or HS models that are poorly calibrated.

Finally, we consider the proxy weighted scoring rule based on Section 7.4. Similar to the previous section, for each simulation, we estimate the proxy average score, denoted by $\bar{R}^*_g = \frac{1}{n}\sum_{t=1}^{n} R_g(F^*_t, L_t)$, and the proxy average benchmark score, denoted by $\bar{R}^*_{g,0} = \frac{1}{n}\sum_{t=1}^{n} R_g(F^*_{t,0}, L_t)$, where $F^*_{t,0}$ is the proxy model estimated using the true realized PIT values and true VaR. The proxy score difference and its sample mean are denoted respectively by $D^*_{g,t} = R_g(F^*_t, L_t) - R_g(F^*_{t,0}, L_t)$ and $\bar{D}^*_g = \frac{1}{n}\sum_{t=1}^{n} D^*_{g,t}$. We estimate the mean and standard deviation of $\bar{D}^*_g$, as well as the DM test rejection rates via simulations. The results are shown in Table 30. The conclusions here are the same as those from Table 27, where the results for the forecast models GARCH.t, GARCH.norm and ARCH.t is very similar to those in Table 29, whereas the HS and EWMA.HS forecast models are penalized heavily (especially in terms of DM rejection rates) as the proxy model structure that we use cannot capture the HS structure properly.

Result Type	$\widehat{F}_t$ | weight	Uniform	Linear	Exponential
Mean of $\bar{D}_g$	GARCH.t	0.0006 (1)	0.0005 (1)	0.0004 (1)
	GARCH.norm	0.0014 (2)	0.0016 (2)	0.0018 (3)
	ARCH.t	0.0151 (5)	0.0123 (5)	0.0093 (5)
	EWMA.HS.500	0.0017 (3)	0.0016 (3)	0.0015 (2)
	HS.500	0.0223 (7)	0.0181 (7)	0.0137 (7)
	EWMA.HS.250	0.0026 (4)	0.0025 (4)	0.0024 (4)
	HS.250	0.0182 (6)	0.0150 (6)	0.0118 (6)
Standard Deviation of $\bar{D}_g$	GARCH.t	0.0258 (1)	0.0239 (1)	0.0216 (1)
	GARCH.norm	0.0432 (3)	0.0480 (4)	0.0533 (4)
	ARCH.t	0.1559 (5)	0.1352 (5)	0.1101 (5)
	EWMA.HS.500	0.0403 (2)	0.0366 (2)	0.0327 (2)
	HS.500	0.1870 (7)	0.1650 (7)	0.1410 (7)
	EWMA.HS.250	0.0476 (4)	0.0444 (3)	0.0421 (3)
	HS.250	0.1685 (6)	0.1520 (6)	0.1366 (6)
DM Rejection rate	GARCH.t	15.8 (1)	14.9 (1)	14.4 (1)
	GARCH.norm	16.4 (2)	15.2 (2)	14.8 (2)
	ARCH.t	83.0 (5)	79.6 (5)	74.8 (5)
	EWMA.HS.500	37.2 (3)	41.9 (3)	46.7 (3)
	HS.500	88.0 (6)	85.0 (6)	81.2 (7)
	EWMA.HS.250	51.7 (4)	57.1 (4)	57.5 (4)
	HS.250	88.6 (7)	85.0 (6)	79.6 (6)

Table 29: The estimated mean and standard deviation of $\bar{D}_g$, and the Diebold Mariano test rejection rate, at varying weight functions. The DGP is based on a GARCH process with Student-t innovations. Results are in % and obtained using 1000 simulations. The values in the brackets are the rankings in ascending order.

Result Type	$\widehat{F}_t$ | weight	Uniform	Linear	Exponential
Mean of $\bar{D}_g^*$	GARCH.t	0.0006 (1)	0.0005 (1)	0.0004 (1)
	GARCH.norm	0.0014 (2)	0.0016 (2)	0.0019 (2)
	ARCH.t	0.0151 (5)	0.0123 (5)	0.0093 (5)
	EWMA.HS.500	0.0036 (3)	0.0041 (3)	0.0045 (3)
	HS.500	0.0246 (7)	0.0218 (7)	0.0190 (7)
	EWMA.HS.250	0.0039 (4)	0.0045 (4)	0.0051 (4)
	HS.250	0.0207 (6)	0.0191 (6)	0.0176 (6)
Standard Deviation of $\bar{D}_g^*$	GARCH.t	0.0258 (1)	0.0239 (1)	0.0216 (1)
	GARCH.norm	0.0435 (2)	0.0483 (2)	0.0535 (2)
	ARCH.t	0.1559 (5)	0.1352 (5)	0.1101 (5)
	EWMA.HS.500	0.0617 (4)	0.0629 (4)	0.0676 (4)
	HS.500	0.2029 (7)	0.1936 (7)	0.1856 (6)
	EWMA.HS.250	0.0548 (3)	0.0570 (3)	0.0631 (3)
	HS.250	0.1902 (6)	0.1871 (6)	0.1863 (7)
DM Rejection rate	GARCH.t	15.8 (1)	14.9 (1)	14.4 (1)
	GARCH.norm	17.8 (2)	16.3 (2)	15.4 (2)
	ARCH.t	83.0 (5)	79.6 (4)	74.8 (3)
	EWMA.HS.500	70.6 (3)	78.5 (3)	81.0 (4)
	HS.500	93.1 (7)	92.4 (6)	91.4 (6)
	EWMA.HS.250	76.8 (4)	85.2 (5)	89.2 (5)
	HS.250	91.8 (6)	92.4 (6)	92.3 (7)

Table 30: The estimated mean and standard deviation of $\bar{D}_g^*$, and the Diebold Mariano test rejection rate, at varying weight functions, based on the weighted scoring rule when we replace the forecast model $\widehat{F}_t$ with the proxy model F_t^* estimated using available data sets. The DGP is based on a GARCH process with Student-t innovations. Results are in % and obtained using 1000 simulations. The values in the brackets are the rankings in ascending order.

Chapter 8 Summary

We have summarized the general concepts of hypothesis testing in Chapter 2, which laid the groundwork to better understand the backtests described in Chapter 3 and Chapter 5. The main contribution in Chapter 3 is the introduction of the framework for testing weighted transformations of the realized PIT values, which nests many of the traditional VaR and realized PIT tests in existing literature. The key advantage of the framework is that it is very flexible as it allows the testers to assign weights that reflect their risk management objectives. We have also introduced the probitnormal score test, which is closely related to the Berkowitz (2001) test, but is a much more flexible test since it can be easily extended to explicitly test for serial independence.

In Chapter 4, we conduct simulation studies on the tests described in Chapter 3. The key findings are that the multinomial tests, which are the extensions of the binomial tests, are more powerful in detecting misspecified forecast distributions. For the spectral tests, since the misspecified forecast distributions that we have chosen have increasing VaR exception rate bias towards the tail of the distribution, spectral tests which place more weight towards the tail are more powerful. The bispectral tests are more powerful than the spectral tests as they contain correlation information between two spectral tests. Effectively, they test for more moments in the tail. We also find that exception-based tests have great difficulty in detecting poorly calibrated HS models, despite of their large VaR bias in the tail of the distribution. This observation is consistent with the findings in Pritsker (2006).

A further advantage of the general framework introduced in Chapter 3 is that it can be easily be extended to allow for explicit testing for serial independence. We have shown how to do this in Chapter 5. In particular, we have introduced the block tests and the martingale difference (MD) tests. Along with the Nass-type size correction, the MD tests serve as a powerful tool to test for misspecification in the dynamics of the forecast distributions. We have conducted simulation studies to understand the size of these extensions in Chapter 6, and the power against different forms of dynamic misspecification.

Finally, some existing and new ideas based on elicitability theory are explored in

Chapter 7. We find that the weighted scoring rule that we consider, which is similar to the weighted CRPS score of Gneiting & Ranjan (2011), produces more reliable Diebold & Mariano (1995) test results. Since the weighted scoring rule is sensitive to VaR bias, these tests are useful to complement the exception-based tests to better detect poorly calibrated HS models.

Although we have developed the framework to cater for different risk management objectives, we have made no specific recommendations as to how to use the framework. A general rule of thumb is that the choice of weight functions should reflect the risk management objectives. For example, for Basel III, where the risk measure of interest is the VaR at the 97.5% and 99% levels, and ES at the 97.5% level, a bispectral test with decreasing/increasing weight functions in the range $(0.975, 0.9995)$ may work well. See Gordy et al. (2017) for more details. For Solvency II, where the risk measure of interest is the one-year VaR at the 99.5% level, we recommend the probitnormal score test in the range $(0.985, 0.995)$. This test have better size performance compared to the binomial test at the 99.5% level, while still placing an adequately large weight at the 99.5% level. The trade-off of specificity to the risk objective for a better size performance is important in this case as the sample size of the data-sets available are usually very small due to the use of one-year VaR.

For the dynamic tests in Chapter 5, based on the results from the simulation studies in Chapter 6, the broad recommendation is to use a one-week lag period (setting $h = 4$ for the conditional tests or $B = 5$ for the block tests), and using the factor described in (5.26) with $\beta_1 = 0.0005$ and $\beta_2 = 0.9995$ for the conditional tests. However, if our goal is to improve an existing model, we may wish to have a test that is as powerful as possible in detecting poor models. One way to do this may be to automate the choice of the lag period based on past ACF information.

Possible future work includes extending the framework to a multivariate setting, which may be useful for banks that want to backtest outputs of multiple trading desks simultaneously. The multivariate framework may also be useful for Solvency II, since one way to compensate for small data-sets is to test the model on multiple similar data-sets. While the Bonferroni test is straightforward to implement, the bounds are too loose when the multiple data-sets are highly correlated, leading to low power of the tests. Hence, a multivariate framework which allow the user to

specify a correlation structure will be particularly useful.

From the results in Chapter 7, we found that tests based on elicitability theory complement the exception-based tests well. It would be interesting to be able to conduct both tests simultaneously while controlling family-wise error. A possible starting point to allow for the complex correlation structure between multiple test statistics would be to use bootstrap test statistics similar to the ideas found in, for example, White (2000) and Hansen (2005).

Finally, financial institutions are usually interested in detecting model failure as soon as possible. Another possible future work would be to apply the framework in a statistical process control setting based on the work of, for example, Hawkins & Zamba (2005).

References

Acerbi, C. (2002). Spectral measures of risk: a coherent represeentation of subjective risk aversion. *J. Banking Finance* **26**, 1505–1518.

Acerbi, C. & Szekely, B. (2014). Backtesting expected shortfall. *MSCI Inc.* .

Acerbi, C. & Székely, B. (2016). L'ES est mort, vive l'ES! Talk at ETH Zurich, available online.

Acerbi, C. & Tasche, D. (2002). On the coherence of expected shortfall. *J. Banking Finance* **26**, 1487–1503.

Artzner, P., Delbaen, F., Eber, J. & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance* **9**, 203–228.

Barone-Adesi, G., Bourgoin, F. & Giannopoulos, K. (1998). Don't look back. *Risk* **11**, 100–103.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society A* **160**, 268–282.

Basel Committee on Banking Supervision (2013). Fundamental review of the trading book: A revised market risk framework. Bank of International Settlements.

Basel Committee on Banking Supervision (2016). Minimum capital requirements for market risk. Bank of International Settlements.

Bellini, F. & Bignozzi, V. (2015). On elicitable risk measures. *Quantitative Finance* **15**, 725–733.

Berkowitz, J. (2001). Testing the accuracy of density forecasts, applications to risk management. *J. Bus. Econ. Statist.* **19**, 465–474.

Berkowitz, J., Christoffersen, P. & Pelletier, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science* **57**, 2213–2227.

Blum, P. (2004). *On some mathematical aspects of dynamic financial analysis*. Ph.D. thesis, ETH Zürich (Swiss Federal Institute of Technology, Zurich).

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, 3–62.

Box, G. & Pierce, D. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J. Amer. Statist. Assoc.* **65**, 1509–1526.

Brockwell, P. & Davis, R. (1991). *Time Series: Theory and Methods*. Springer, New York, 2nd ed.

Brockwell, P. & Davis, R. (2003). Introduction to time series and forecasting. *Springer* .

Cai, Y. & Krishnamoorthy, K. (2006). Exact size and power properties of five tests for multinomial proportions. *Communications in Statistic, Simulation and Computation* **35**, 149–160.

Casella, G. & Berger, R. (2001). *Statistical Inference*. Cengage Learning; 2nd edition.

Chen, L. & Shi, J. (2011). Empirical likelihood hypothesis test on mean with inequality constraints. *Science China Mathematics* **9**, 1847–1857.

Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review* **39**.

Christoffersen, P. & Pelletier, D. (2004). Backtesting Value-at-Risk: a duration-based approach. *Journal of Econometrics* **2**, 84–108.

Clift, S., Costanzino, N. & Curran, M. (2015). Empirical performance of backtesting methods for expected shortfall. Working paper.

Costanzino, N. & Curran, M. (2015). Backtesting general spectral risk measures with application to expected shortfall. *Working paper.* .

Costanzino, N. & Curran, M. (2016). A simple traffic light approach to backtesting expected shortfall. *Working paper* .

Davé, R. & Stahl, G. (1998). On the accuracy of var estimates based on the variance-covariance approach. *Risk Measurement, Econometrics and Neural Networks* , 198–232.

Davis, M. (2013). Consistency of risk measure estimates. Working paper.

Davis, M. (2016). Verification of internal risk measure estimates. *Statistics & Risk Modeling* .

Diebold, F., Gunther, T. & Tay, A. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39**, 863–883.

Diebold, F., Hahn, J. & Tay, A. (1999). Multivariate density forecasts and calibration in financial risk management: High-frequency returns on foreign exchange. *The Review of Economics and Statistics* **81**, 661–673.

Diebold, F. & Mariano, R. (1995). Comparing predictive accuracy. *Journal of business & economic statistics* **13**, 253–263.

Dowd, K. (2008). A moments-based procedure for evaluating risk forecasting models. *The Analytics of Risk Model Validation* , 45–58.

Dumitrescu, E. I., Hurlin, C. & Pham, V. (2012). Backtesting value-at-risk: from dynamic quantile to dynamic binary tests. *Preprint.* .

Ehm, W., Gneiting, T., Jordan, A. & Kruger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, choquet representations, and forecast rankings. *Journal of the Royal Statistical Society* **78**, 505–562.

Emmer, S., Kratz, M. & Tasche, D. (2015). What is the best risk measure in practice? *Journal of Risk* **18**, 31–60.

Engle, R. & Manganelli, S. (2004). Caviar: conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics* **22**, 367–381.

Fernandez, C. & Steel, M. (1998). On bayesian modelling of fat tails and skewness. *Journal of the American Statistical Association* **93**, 359–371.

Fissler, T. & Ziegel, F. (2015). Higher order elicitability and osband's principle. *The Annals of Statistics 2016* **44**, 1680–1707.

Föllmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and Stochastics* **6**, 429–447.

Föllmer, H. & Schied, A. (2011). *Stochastic Finance An Introduction in Discrete Time*. Berlin New York: Walter de Gruyter, 3rd ed.

Fraga, A., Gomes, M. & de Haan, L. (2003). A new class of semiparametric estimators of the second order parameter. *Portugaliae Mathematica* **60**, 194–213.

Giacomini, R. & White, H. (2006). Tests of conditional predictive ability. *Econometrica* **74**, 1545–1578.

Glosten, L., Jagannathan, R. & Runkle, D. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *J. Finance* **48**, 1779–1801.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**, 746–762.

Gneiting, T. & Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Gneiting, T. & Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics* **29**, 411–422.

Gomes, M., I. & Martins, M. (2002). Asymptotically unbiased estimators of the tail index based on external estimation of the second order parameter. *Extremes* **5**, 5–31.

Gomes, M., I. & Pestana, D. (2007). A sturdy reduced-bias extreme quantile (var) estimator. *Journal of the American Statistical Association* **102**, 280–292.

Gordy, M., Lok, H. & McNeil, A. (2017). Spectral backtests of forecast distributions with application to risk management. *Working Paper* .

Hansen, P. (2005). A test for superior predictive ability. *Journal of business & economic statistics* **23**, 365–380.

Harvey, D., Leybourne, S. & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting* **13**, 281–291.

Hawkins, D. & Zamba, K. (2005). Statistical process control for shifts in mean or variance using a change point formulation. *Technometrics* **47**, 164–173.

Heyde, C., Kou, S. & Peng, X. (2007). What is a good risk measure: bridging the gaps between data, coherent risk measures, and insurance risk measures. *Preprint, Columbia University* .

Hommel, G., Bretz, F. & Maurer, W. (2011). Multiple hypotheses testing based on ordered p values - a historical survey with applications to medical research. *Journal of Biopharmaceutical Statistics* **4**, 595–609.

Hull, J. & White, A. (1998). Incorporating volatility updating into the historical simulation method for value-at-risk. *J. Risk* **1**, 5–19.

Jarvis, S., Sharpe, J. & Smith, A. (2016). Ersatz model tests. *SSRN working paper* .

Kauppi, H. & Saikkonen, P. (2008). Predicting U.S. recessions with dynamic binary response models. *The Review of Economics and Statistics* **90**, 777–791.

Kerkhof, J. & Melenberg, B. (2004). Backtesting for risk-based regulatory capital. *Journal of Banking and Finance* **28**, 1845–1865.

Kratz, M., Lok, H. & McNeil, A. (2016). A multinomial test to discriminate between models. *Proceedings of ASTIN conference 2016* , 1–9.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives* **3**, 73–84.

Lambert, N., Pennock, D. & Shoham, Y. (2008). Eliciting properties of probability distributions. *Proceedings of the 9th ACM Conference on Electronic Commerce* .

Ljung, G. & Box, G. (1978). On a measure of lack of fit in time series models. *Biometrika* **65**, 297–303.

Marshall, C. & Siegel, M. (1997). Value at risk: Implementing a risk measurement standard. *Journal of Derivatives* **4**, 91–110.

McNeil, A. & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *J. Empirical Finance* **7**, 271–300.

McNeil, A. J., Frey, R. & Embrechts, P. (2015). Quantitative risk management: Concepts, techniques and tools, second edition. *Princeton University Press, Princeton* .

Nass, C. (1959). The χ^2-test for small expectations in contingency tables, with special reference to accidents and absenteeism. *Biometrika* **46**, 365–385.

Nolde, N. & Ziegel, J. (2016). Elicitability and backtesting. *ArXiv e-prints* .

O'Brien, J. & Szerszen, P. (2014). An evaluation of bank VaR measures for market risk during and before the financial crisis. Tech. Rep. 2014-21, Federal Reserve Board, Wahington, D.C. Finance and Economics Discussion Series, Divisons of Research & Statistics and Monetary Affairs.

Osband, K. & Reichelstein, S. (1985). Information-eliciting compensation schemes. *Journal of Public Economics* **27**, 107–115.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Mag. 5th Ser.* **50**, 157–175.

Pearson, K. (1932). Experimental discussion of the (χ^2,p) test for goodness of fit. *Biometrika* **24**, 351–381.

Pérignon, C. & Smith, D. (2010). The level and quality of Value-at-Risk disclosure by commercial banks. *Journal of Banking and Finance* **34**, 362–377.

Pritsker, M. (2006). The hidden dangers of historical simulation. *Journal of Banking and Finance* **30**, 561–582.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics* **23**, 470–472.

Saerens, M. (2000). Building cost functions minimizing to some summary statistics. *IEEE Transactions on Neural Networks* **11**, 1263–1271.

Savage, L. (1971). Elicitation of personal probabilities and expectations. *Journal of hte American Statistical Association* **66**, 783–801.

Tasche, D. (2002). Expected shortfall and beyond. *J. Banking Finance* **26**, 1519–1533.

Thomson, W. (1979). Eliciting production possibilities from a well-informed manager,. *Journal of Economic Theory* **20**, 360–480.

Vessereau, A. (1958). Sur les conditions d'application du critérium χ^2 de pearson. *Revue de Statistique Appliquée* **6**, 83–96.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.* **73**, 812–815–441.

West, K. (1996). Asymptotic inference about predictive ability. *Econometrica* **64**, 1067–1084.

White, H. (2000). A reality check for data snooping. *Econometrica* **68**, 1097–1126.

Ziegel, J. (2013). Coherence and elicitability. Preprint.

Zumbach, G. (2006). Backtesting risk methodologies from one day to one year. *Journal of Risk* **9**, 55–91.

Appendix A Fisher information matrix for truncated probitnormal score test

The following identities are useful for dealing with the probitnormal distribution:

$$\int_{\alpha_1}^{\alpha_2} \Phi^{-1}(u)\mathrm{d}u = \phi(\Phi^{-1}(\alpha_1)) - \phi(\Phi^{-1}(\alpha_2)) \tag{A.1}$$

$$\int_{\alpha_1}^{\alpha_2} \left(\Phi^{-1}(u)^2 - 1\right) \mathrm{d}u = \Phi^{-1}(\alpha_1)\phi(\Phi^{-1}(\alpha_1)) - \Phi^{-1}(\alpha_2)\phi(\Phi^{-1}(\alpha_2)). \tag{A.2}$$

Let $\xi(p \mid \boldsymbol{\theta}) = (\Phi^{-1}(p) - \mu)/\sigma$

$$-\frac{\partial^2}{\partial\mu^2} \ln L(\boldsymbol{\theta} \mid P_t) = \begin{cases} \frac{\phi(\xi(\alpha_1|\boldsymbol{\theta}))\left(\phi(\xi(\alpha_1|\boldsymbol{\theta}))+\xi(\alpha_1|\boldsymbol{\theta})\Phi(\xi(\alpha_1|\boldsymbol{\theta}))\right)}{\sigma^2\Phi(\xi(\alpha_1|\boldsymbol{\theta}))^2} & P_t \leqslant \alpha_1, \\ \frac{1}{\sigma^2} & \alpha_1 < P_t < \alpha_2, \\ \frac{\phi(\xi(\alpha_2|\boldsymbol{\theta}))\left(\phi(\xi(\alpha_2|\boldsymbol{\theta}))-\xi(\alpha_2|\boldsymbol{\theta})\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))\right)}{\sigma^2\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))^2} & P_t \geqslant \alpha_2. \end{cases} \tag{A.3}$$

$$-\frac{\partial^2}{\partial\sigma^2} \ln L(\boldsymbol{\theta} \mid P_t) = \begin{cases} \frac{\phi(\xi(\alpha_1|\boldsymbol{\theta}))\left(\xi(\alpha_1|\boldsymbol{\theta})^2\phi(\xi(\alpha_1|\boldsymbol{\theta}))+\xi(\alpha_1|\boldsymbol{\theta})^3\Phi(\xi(\alpha_1|\boldsymbol{\theta}))-2\xi(\alpha_1|\boldsymbol{\theta})\Phi(\xi(\alpha_1|\boldsymbol{\theta}))\right)}{\sigma^2\Phi(\xi(\alpha_1|\boldsymbol{\theta}))^2} & P_t \leqslant \alpha_1, \\ \frac{3\xi(P_t|\boldsymbol{\theta})^2-1}{\sigma^2} & \alpha_1 < P_t < \alpha_2, \\ \frac{\phi(\xi(\alpha_2|\boldsymbol{\theta}))\left(\xi(\alpha_2|\boldsymbol{\theta})^2\phi(\xi(\alpha_2|\boldsymbol{\theta}))-\xi(\alpha_2|\boldsymbol{\theta})^3\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))+2\xi(\alpha_2|\boldsymbol{\theta})\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))\right)}{\sigma^2\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))^2} & P_t \geqslant \alpha_2. \end{cases} \tag{A.4}$$

$$-\frac{\partial^2}{\partial\mu\partial\sigma} \ln L(\boldsymbol{\theta} \mid P_t) = \begin{cases} \frac{\phi(\xi(\alpha_1|\boldsymbol{\theta}))\left(\phi(\xi(\alpha_1|\boldsymbol{\theta}))\xi(\alpha_1|\boldsymbol{\theta})-\Phi(\xi(\alpha_1|\boldsymbol{\theta}))+\xi(\alpha_1|\boldsymbol{\theta})^2\Phi(\xi(\alpha_1|\boldsymbol{\theta}))\right)}{\sigma^2\Phi(\xi(\alpha_1|\boldsymbol{\theta}))^2} & P_t \leqslant \alpha_1, \\ \frac{2\xi(P_t|\boldsymbol{\theta})}{\sigma^2} & \alpha_1 < P_t < \alpha_2, \\ \frac{\phi(\xi(\alpha_2|\boldsymbol{\theta}))\left(\phi(\xi(\alpha_2|\boldsymbol{\theta}))\xi(\alpha_2|\boldsymbol{\theta})+\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))-\xi(\alpha_2|\boldsymbol{\theta})^2\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))\right)}{\sigma^2\overline{\Phi}(\xi(\alpha_2|\boldsymbol{\theta}))^2} & P_t \geqslant \alpha_2. \end{cases} \tag{A.5}$$

For $\boldsymbol{\theta}_0 = (0, 1)'$

$$\begin{aligned} I(\boldsymbol{\theta}_0)_{1,1} = {} & \phi(\Phi^{-1}(\alpha_1))^2/\alpha_1 + \phi(\Phi^{-1}(\alpha_2))^2/(1-\alpha_2) \\ & + \phi(\Phi^{-1}(\alpha_1))\Phi^{-1}(\alpha_1) - \phi(\Phi^{-1}(\alpha_2))\Phi^{-1}(\alpha_2) + (\alpha_2 - \alpha_1) \end{aligned} \tag{A.6}$$

$$\begin{aligned} I(\boldsymbol{\theta}_0)_{2,2} = {} & \phi(\Phi^{-1}(\alpha_1))^2\Phi^{-1}(\alpha_1)^2/\alpha_1 + \phi(\Phi^{-1}(\alpha_1))\Phi^{-1}(\alpha_1)^3 \\ & + \phi(\Phi^{-1}(\alpha_1))\Phi^{-1}(\alpha_1) + \phi(\Phi^{-1}(\alpha_2))^2\Phi^{-1}(\alpha_2)^2/(1-\alpha_2) \\ & - \phi(\Phi^{-1}(\alpha_2))\Phi^{-1}(\alpha_2)^3 - \phi(\Phi^{-1}(\alpha_2))\Phi^{-1}(\alpha_2) + 2(\alpha_2 - \alpha_1) \end{aligned} \tag{A.7}$$

$$I(\boldsymbol{\theta}_0)_{1,2} = \phi(\Phi^{-1}(\alpha_1))^2\Phi^{-1}(\alpha_1)/\alpha_1 + \phi(\Phi^{-1}(\alpha_1))\big(1+\Phi^{-1}(\alpha_1)^2\big) + \phi(\Phi^{-1}(\alpha_2))^2\Phi^{-1}(\alpha_2)/(1-\alpha_2) - \phi(\Phi^{-1}(\alpha_2))\big(1+\Phi^{-1}(\alpha_2)^2\big) \quad \text{(A.8)}$$

Appendix B Conditional spectral and bispectral test statistic variance

B.1 Using $f(P_t) = \tilde{W}_{v,t}$ to test for serial independence

We denote

$$\tilde{W}_{v_1,t} = W_{v_1,t} - \mu_{v_1}\,, \tag{B.1}$$

$$\tilde{W}_{v_2,t} = W_{v_2,t} - \mu_{v_2}\,, \tag{B.2}$$

$$\tilde{L}_{k,t} = \tilde{W}_{v_1,t}\,\tilde{W}_{v_1,t-k}\,, \tag{B.3}$$

$$\overline{W}_{v_1} = \frac{1}{n-h}\sum_{t=h+1}^{n}\tilde{W}_{v_1,t}\,, \tag{B.4}$$

$$\overline{W}_{v_2} = \frac{1}{n-h}\sum_{t=h+1}^{n}\tilde{W}_{v_2,t}\,, \tag{B.5}$$

$$\overline{L}_{k} = \frac{1}{n-h}\sum_{t=h+1}^{n}\tilde{L}_{k,t}\,. \tag{B.6}$$

Conditional spectral test. For notational simplicity, we assume $v_1 = v$. It can be shown that

$$\begin{aligned}\frac{\operatorname{var}(S)}{(n-h)^2} &= (\Sigma_V^{-1})_{1,1}^2\,\operatorname{var}(\overline{W}_{v_1}^2) + 2\sum_{k=1}^{h}(\Sigma_V^{-1})_{1,1}\,(\Sigma_V^{-1})_{k+1,k+1}\,\operatorname{cov}(\overline{W}_{v_1}^2,\overline{L}_k^2)\\ &\quad+\sum_{k=1}^{h}(\Sigma_V^{-1})_{k+1,k+1}^2\,\operatorname{var}(\overline{L}_k^2)\\ &\quad+ I_{\{h\geqslant 2\}}\,2\sum_{i=1}^{h-1}\sum_{j=i+1}^{h}(\Sigma_V^{-1})_{i+1,i+1}\,(\Sigma_V^{-1})_{j+1,j+1}\operatorname{cov}(\overline{L}_i^2,\overline{L}_j^2)\,.\end{aligned} \tag{B.7}$$

Conditional bispectral test. It can be shown that

$$\begin{aligned}
\frac{\operatorname{var}(S)}{(n-h)^2} &= (\Sigma_V^{-1})_{1,1}^2 \operatorname{var}(\overline{W}_{v_1}^2) + (\Sigma_V^{-1})_{2,2}^2 \operatorname{var}(\overline{W}_{v_2}^2) + 4\,(\Sigma_V^{-1})_{1,2}^2 \operatorname{var}(\overline{W}_{v_1}\,\overline{W}_{v_2}) \\
&\quad + 2\,(\Sigma_V^{-1})_{1,1}\,(\Sigma_V^{-1})_{2,2} \operatorname{cov}(\overline{W}_{v_1}^2, \overline{W}_{v_2}^2) \\
&\quad + 4\,(\Sigma_V^{-1})_{1,2} \left((\Sigma_V^{-1})_{1,1}\, \operatorname{cov}(\overline{W}_{v_1}^2, \overline{W}_{v_1}\,\overline{W}_{v_2}) + (\Sigma_V^{-1})_{2,2}\, \operatorname{cov}(\overline{W}_{v_2}^2, \overline{W}_{v_1}\,\overline{W}_{v_2}) \right) \\
&\quad + 2 \sum_{k=1}^{h} (\Sigma_V^{-1})_{1,1}\,(\Sigma_V^{-1})_{k+2,k+2}\, \operatorname{cov}(\overline{W}_{v_1}^2, \overline{L}_k^2) \\
&\quad + 2 \sum_{k=1}^{h} (\Sigma_V^{-1})_{2,2}\,(\Sigma_V^{-1})_{k+2,k+2}\, \operatorname{cov}(\overline{W}_{v_2}^2, \overline{L}_k^2) \\
&\quad + 2 \sum_{k=1}^{h} (\Sigma_V^{-1})_{1,2}\,(\Sigma_V^{-1})_{k+2,k+2}\, \operatorname{cov}(\overline{W}_{v_1}\,\overline{W}_{v_2}, \overline{L}_k^2) \\
&\quad + \sum_{k=1}^{h} (\Sigma_V^{-1})_{k+2,k+2}^2\, \operatorname{var}(\overline{L}_k^2) \\
&\quad + I_{\{h \geqslant 2\}}\, 2 \sum_{i=1}^{h-1} \sum_{j=i+1}^{h} (\Sigma_V^{-1})_{i+2,i+2}\,(\Sigma_V^{-1})_{j+2,j+2} \operatorname{cov}(\overline{L}_i^2, \overline{L}_j^2)\,. \qquad \text{(B.8)}
\end{aligned}$$

We can calculate

$$(n-h)^4 \operatorname{var}(\overline{W}_{v_1}^2) = (n-h)(\mathrm{E}\,(\tilde{W}_{v_1,t}^4) - \sigma_{v_1}^4) + 2(n-h)(n-h-1)\sigma_{v_1}^4\,, \tag{B.9}$$

$$(n-h)^4 \operatorname{var}(\overline{W}_{v_2}^2) = (n-h)(\mathrm{E}\,(\tilde{W}_{v_2,t}^4) - \sigma_{v_2}^4) + 2(n-h)(n-h-1)\sigma_{v_2}^4\,, \tag{B.10}$$

$$\begin{aligned}(n-h)^4 \operatorname{var}(\overline{W}_{v_1}\,\overline{W}_{v_2}) &= (n-h)(\mathrm{E}\,(\tilde{W}_{v_1,t}^2\tilde{W}_{v_2,t}^2) - \sigma_{v_1,g_2}^2) \\ &\quad + (n-h)(n-h-1)(\sigma_{v_1}^2\sigma_{v_2}^2 + \sigma_{v_1,v_2}^2)\,,\end{aligned} \tag{B.11}$$

$$\begin{aligned}(n-h)^4 \operatorname{cov}(\overline{W}_{v_1}^2, \overline{W}_{v_2}^2) &= (n-h)(\mathrm{E}\,(\tilde{W}_{v_1,t}^2\tilde{W}_{v_2,t}^2) - \sigma_{v_1}^2\sigma_{v_2}^2) \\ &\quad + 2(n-h)(n-h-1)\sigma_{v_1,v_2}^2\,,\end{aligned} \tag{B.12}$$

$$\begin{aligned}(n-h)^4 \operatorname{cov}(\overline{W}_{v_1}^2, \overline{W}_{v_1}\,\overline{W}_{v_2}) &= (n-h)(\mathrm{E}\,(\tilde{W}_{v_1,t}^3\tilde{W}_{v_2,t}) - \sigma_{v_1}^2\sigma_{v_1,v_2}) \\ &\quad + 2(n-h)(n-h-1)\sigma_{v_1}^2\sigma_{v_1,v_2}\,,\end{aligned} \tag{B.13}$$

$$\begin{aligned}(n-h)^4 \operatorname{cov}(\overline{W}_{v_2}^2, \overline{W}_{v_1}\,\overline{W}_{v_2}) &= (n-h)(\mathrm{E}\,(\tilde{W}_{v_1,t}\tilde{W}_{v_2,t}^3) - \sigma_{v_2}^2\sigma_{v_1,v_2}) \\ &\quad + 2(n-h)(n-h-1)\sigma_{v_2}^2\sigma_{v_1,v_2}\,,\end{aligned} \tag{B.14}$$

$$\begin{aligned}(n-h)^4 \operatorname{cov}(\overline{W}_{v_1}^2, \overline{L}_k^2) &= (2(n-h-k)+k)(\mathrm{E}\,(\tilde{W}_{v_1,t}^4)\sigma_{v_1}^2 - \sigma_{v_1}^6) \\ &\quad + 2(n-h-k)(\mathrm{E}\,(\tilde{W}_{v_1,t}^3)^2) \\ &\quad + 4\,\max(n-h-2k,0)\,\sigma_{v_1}^6\,,\end{aligned} \tag{B.15}$$

$$\begin{aligned}(n-h)^4 \operatorname{cov}(\overline{W}_{v_2}^2, \overline{L}_k^2) &= (2(n-h-k)+k)(\mathrm{E}\,(\tilde{W}_{v_1,t}^2\tilde{W}_{v_2,t}^2)\sigma_{v_1}^2 - \sigma_{v_1}^4\sigma_{v_2}^2) \\ &\quad + 2(n-h-k)(\mathrm{E}\,(\tilde{W}_{v_1,t}^2\tilde{W}_{v_2,t}^2) \\ &\quad + 4\,\max(n-h-2k,0)\,\sigma_{v_1}^2\sigma_{v_1,v_2}^2\,,\end{aligned} \tag{B.16}$$

$$\begin{aligned}(n-h)^4 \operatorname{cov}(\overline{W}_{v_1}\,\overline{W}_{v_2}, \overline{L}_k^2) &= (2(n-h-k)+k)(\mathrm{E}\,(\tilde{W}_{v_1,t}^3\tilde{W}_{v_2,t})\sigma_{v_1}^2 - \sigma_{v_1}^4\sigma_{v_1,v_2}) \\ &\quad + 2(n-h-k)\mathrm{E}\,(\tilde{W}_{v_1,t}^3)\mathrm{E}\,(\tilde{W}_{v_1,t}^2\tilde{W}_{v_2,t}) \\ &\quad + 4\,\max(n-h-2k,0)\,\sigma_{v_1}^4\sigma_{v_1,v_2}\,,\end{aligned} \tag{B.17}$$

$$\begin{aligned}(n-h)^4 \operatorname{var}(\overline{L}_k^2) &= (n-h)(\mathrm{E}\,(\tilde{W}_{v_1,t}^4)^2 - \sigma_{v_1}^8) \\ &\quad + 4\,(n-h-k)\,\mathrm{E}\,(\tilde{W}_{v_1,t}^4)\sigma_{v_1}^4 \\ &\quad + (2\,(n-h)(n-h-1) - 4\,(n-h-k))\sigma_{v_1}^8 \\ &\quad + 2\,(n-h-k)(\mathrm{E}\,(\tilde{W}_{v_1,t}^4)\sigma_{v_1}^4 - \sigma_{v_1}^8)\,,\end{aligned} \tag{B.18}$$

$$\begin{aligned}(n-h)^4 \operatorname{cov}(\overline{L}_i^2, \overline{L}_j^2 \mid j>i) &= (6i + 4((n-h)-(i+j)) + 2(j-i))(\mathrm{E}\,(\tilde{W}_{v_1,t}^4)\sigma_{v_1}^4 - \sigma_{v_1}^8) \\ &\quad + I_{\{j=2i\}}(n-h-i)\mathrm{E}\,(\tilde{W}_{v_1,t}^3)^2\sigma_{v_1}^2 \\ &\quad + (n-h-j)\sigma_{v_1}^8\,.\end{aligned} \tag{B.19}$$

Conditional binomial. This is a special case of the conditional spectral test, where $\tilde{W}_{v_1,t} = I_{\{P_t > \alpha\}} - (1-\alpha)$. We can easily calculate

$$\mathrm{E}\,(\tilde{W}^3_{v_1,t}) = \alpha(1-\alpha)(\alpha^2 - (1-\alpha)^2)\,, \tag{B.20}$$

$$\mathrm{E}\,(\tilde{W}^4_{v_1,t}) = \alpha(1-\alpha)(\alpha^3 + (1-\alpha)^3)\,, \tag{B.21}$$

using the fact that $\tilde{W}_{v_1,t}$ take values α with probability $(1-\alpha)$, and $(\alpha - 1)$ with probability α.

Conditional spectral and bispectral. In the continuous weighting case when $\mathrm{d}v_i(u) = g_i(u)\,\mathrm{d}u$ with g_i satisfying Assumption 3.1 for $i = 1, 2$, we can computed $\mathrm{E}\,(\tilde{W}^p_{v_1,t})$, $\mathrm{E}\,(\tilde{W}^p_{v_2,t})$ and $\mathrm{E}\,(\tilde{W}^{p_1}_{v_1,t}\,\tilde{W}^{p_2}_{v_2,t})$ using the results in Lemma 3.4.

Conditional PNS. In the case when v_1 and v_2 are the PNS weight measures in (3.80) and (3.81), we can computed $\mathrm{E}\,(\tilde{W}^p_{v_1,t})$, $\mathrm{E}\,(\tilde{W}^p_{v_2,t})$ and $\mathrm{E}\,(\tilde{W}^{p_1}_{v_1,t}\,\tilde{W}^{p_2}_{v_2,t})$ using the results in Lemma 3.6.

B.2 Using a generic factor $f(P_t) - \mu_f$ to test for serial independence

We denote $X_t = f(P_t)$, $\tilde{X}_t = X_t - \mu_X$ and $\sigma_X^2 = \mathrm{var}(\tilde{X}_t)$, where $\mu_f = \mathrm{E}\,(X_t)$ and σ_X^2 are both evaluated under the assumption that (P_t) are iid uniform. The equations are similar to those in Appendix B.1, except that we replace $\tilde{L}_{k,t}$ with

$$\tilde{L}_{k,t} = \tilde{W}_{v_1,t}\,\tilde{X}_{t-k}\,, \tag{B.22}$$

and change the following equations accordingly:

$$\begin{aligned}(n-h)^4\ \mathrm{cov}(\overline{W}_{v_1}^2, \overline{L}_k^2) = {}& \big(n-h-k\big)\big(\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_1,t}^2)\sigma_{v_1}^2 - \sigma_{v_1}^4\sigma_X^2\big)\\ &+ \big(n-h\big)\big(\mathrm{E}\,(\tilde{W}_{v_1,t}^4)\sigma_X^2 - \sigma_{v_1}^4\sigma_X^2\big)\\ &+ 2(n-h-k)\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_1,t})\,\mathrm{E}\,(\tilde{W}_{v_1,t}^3)\\ &+ 4\ \max(n-h-2k,0)\,\sigma_{v_1}^2\,\mathrm{E}\,(\tilde{X}_t\,\tilde{W}_{v_1,t})^2\,,\end{aligned} \tag{B.23}$$

$$\begin{aligned}(n-h)^4\ \mathrm{cov}(\overline{W}_{v_2}^2, \overline{L}_k^2) = {}& \big(n-h-k\big)\big(\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_2,t}^2)\sigma_{v_1}^2 - \sigma_{v_1}^2\sigma_{v_2}^2\sigma_X^2\big)\\ &+ \big(n-h\big)\big(\mathrm{E}\,(\tilde{W}_{v_1,t}^2\,\tilde{W}_{v_2,t}^2)\sigma_X^2 - \sigma_{v_1}^2\sigma_{v_2}^2\sigma_X^2\big)\\ &+ 2(n-h-k)\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_2,t})\,\mathrm{E}\,(\tilde{W}_{v_1,t}^2\,\tilde{W}_{v_2,t})\\ &+ 4\ \max(n-h-2k,0)\,\mathrm{E}\,(\tilde{X}_t\,\tilde{W}_{v_1,t})\,\mathrm{E}\,(\tilde{X}_t\,\tilde{W}_{v_2,t})\,\mathrm{E}\,(\tilde{W}_{v_1,t}\,\tilde{W}_{v_2,t})\,,\end{aligned} \tag{B.24}$$

$$\begin{aligned}(n-h)^4\ \mathrm{cov}(\overline{W}_{v_1}\,\overline{W}_{v_2}, \overline{L}_k^2) = {}& \big(n-h-k\big)\big(\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_1,t}\,\tilde{W}_{v_2,t})\sigma_{v_1}^2 - \sigma_{v_1,v_2}\sigma_{v_1}^2\sigma_X^2\big)\\ &+ \big(n-h\big)\big(\mathrm{E}\,(\tilde{W}_{v_1,t}^3\,\tilde{W}_{v_2,t})\sigma_X^2 - \sigma_{v_1,v_2}\sigma_{v_1}^2\sigma_X^2\big)\\ &+ (n-h-k)\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_1,t})\,\mathrm{E}\,(\tilde{W}_{v_1,t}^2\,\tilde{W}_{v_2,t})\\ &+ (n-h-k)\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_2,t})\,\mathrm{E}\,(\tilde{W}_{v_1,t}^3)\\ &+ 2\,\sigma_{v_1,v_2}\mathrm{E}\,(\tilde{X}_t\,\tilde{W}_{v_1,t})\,\max(n-h-2k,0)\big(\mathrm{E}\,(\tilde{X}_t\,\tilde{W}_{v_1,t}) + \sigma_{v_1}^2\big)\,,\end{aligned} \tag{B.25}$$

$$\begin{aligned}(n-h)^4\ \mathrm{var}(\overline{L}_k^2) = {}& \big(n-h\big)\big(\mathrm{E}\,(\tilde{W}_{v_1,t}^4)\,\mathrm{E}\,(\tilde{X}_t^4) - \sigma_{v_1}^4\sigma_X^4\big)\\ &+ 4\,(n-h-k)\,\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_1,t}^2)\,\sigma_{v_1}^2\sigma_X^2\\ &+ \big(2\,(n-h)(n-h-1) - 4\,(n-h-k)\big)\sigma_{v_1}^4\sigma_X^4\\ &+ 2\,\big(n-h-k\big)\big(\mathrm{E}\,(\tilde{X}_t^4)\sigma_{v_1}^4 - \sigma_{v_1}^4\sigma_X^4\big)\,,\end{aligned} \tag{B.26}$$

and

$$
\begin{aligned}
(n-h)^4\,\mathrm{cov}(\overline{L}_i^2, \overline{L}_j^2 \mid j>i) = {} & 2i(A+B+C) \\
& + \big((n-h)-(i+j)\big)(A+B+2C) \\
& + \big(j-i\big)(A+C) \\
& + I_{\{j=2i\}}(n-h-i)\mathrm{E}\,(\tilde{X}_t^3)\mathrm{E}\,(\tilde{W}_{v_1,t}^3)\mathrm{E}\,(\tilde{X}_t\,\tilde{W}_{v_1,t}) \\
& + (n-h-j)\mathrm{E}\,(\tilde{X}_t\,\tilde{W}_{v_1,t})^2\,\sigma_{v_1}^2\sigma_X^2\,, \qquad \text{(B.27)}
\end{aligned}
$$

where we denote

$$
\begin{aligned}
A &= \mathrm{E}\,(\tilde{W}_{v_1,t}^4)\,\sigma_X^4 - \sigma_{v_1}^4\sigma_X^4\,, && \text{(B.28)}\\
B &= \mathrm{E}\,(\tilde{X}_t^4)\,\sigma_{v_1}^4 - \sigma_{v_1}^4\sigma_X^4\,, && \text{(B.29)}\\
C &= \mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_1,t}^2)\,\sigma_{v_1}^2\sigma_X^2 - \sigma_{v_1}^4\sigma_X^4\,. && \text{(B.30)}
\end{aligned}
$$

In the case where $X_t = f^*(\tilde{P}_t)$, where $f^*(\tilde{P}_t)$ is defined by (5.26) in Section 5.2.4, we can calculate $\mu_X = \big[-\frac{1}{2}(1-u)^2\big]_{u=\beta_1}^{\beta_2}$ based on the results from Section 3.4.3, and $\mathrm{E}\,(\tilde{X}_t^p)$ for $p \geqslant 2$ can be calculated using the results in Lemma 3.4. We also need to calculate $\mathrm{E}\,(\tilde{X}_t\,\tilde{W}_{v_i,t})$, $\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_i,t})$, $\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_i,t}^2)$ and $\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_1,t}\,\tilde{W}_{v_2,t})$.

Recall from (5.27) that we can write

$$
X_t = 2(Y_{1,t}+Y_{2,t})\,, \quad \text{where } Y_{1,t} = \int_{\gamma_1}^{\gamma_2} I_{\{P_t>u\}}\mathrm{d}u\,, \text{ and } Y_{2,t} = \int_{\gamma_1}^{\gamma_2} I_{\{1-P_t>u\}}\mathrm{d}u\,, \qquad \text{(B.31)}
$$

with $\gamma_1 = \frac{1}{2}(1+\beta_1)$, $\gamma_2 = \frac{1}{2}(1+\beta_2)$, and we chose β_1 and β_2 such that $\gamma_1 \leqslant \alpha_1$ and $\gamma_2 \geqslant \alpha_2$. Using the above relation, we can calculate

$$
\begin{aligned}
\mu_Y = \mathrm{E}\,(Y_{1,t}) = \mathrm{E}\,(Y_{2,t}) &= \frac{1}{4}\mu_X\,, && \text{(B.32)}\\
\sigma_Y^2 = \mathrm{var}(Y_{1,t}) = \mathrm{var}(Y_{2,t}) &= \frac{1}{8}\sigma_X^2 + \mu_Y^2 && \text{(B.33)}\\
\mu_{Y,2} = \mathrm{E}\,(Y_{1,t}^2) = \mathrm{E}\,(Y_{2,t}^2) &= \sigma_Y^2 + \mu_Y^2\,. && \text{(B.34)}
\end{aligned}
$$

We denote $\tilde{Y}_{1,t} = Y_{1,t} - \mu_Y$ and $\tilde{Y}_{2,t} = Y_{2,t} - \mu_Y$. Using the results

$$
\begin{aligned}
\mathrm{E}\,(\tilde{Y}_{2,t}\,W_{v,t}) = \mathrm{E}\,(\tilde{Y}_{2,t}\tilde{W}_{v,t}) &= -\mu_v\,\mu_Y\,, && \text{(B.35)}\\
\mathrm{E}\,(\tilde{Y}_{2,t}^2\,W_{v,t}) &= \mu_v\,\mu_Y^2\,, && \text{(B.36)}\\
\mathrm{E}\,(\tilde{Y}_{2,t}^2\,\tilde{W}_{v,t}) &= \mu_v(2\mu_Y^2 - \mu_{Y,2})\,, && \text{(B.37)}\\
\mathrm{E}\,(\tilde{Y}_{1,t}\,\tilde{Y}_{2,t}\,\tilde{W}_{v,t}) &= \mu_Y(\mu_v\,\mu_Y - \mathrm{E}\,(\tilde{Y}_{1,t}\,\tilde{W}_{v,t}))\,, && \text{(B.38)}
\end{aligned}
$$

for $i = 1, 2$, we can calculate

$$\mathrm{E}\,(\tilde{X}_t\,\tilde{W}_{v_i,t}) = 2(\mathrm{E}\,(\tilde{Y}_{1,t}\,\tilde{W}_{v_i,t}) - \mu_{v_i}\,\mu_Y)\,, \tag{B.39}$$

$$\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_i,t}) = 4\left(\mathrm{E}\,(\tilde{Y}_{1,t}^2\,\tilde{W}_{v_i,t}) + \mu_{v_i}(2\mu_Y^2 - \mu_{Y,2}) + 2\mu_Y\big(\mu_{v_i}\,\mu_Y - \mathrm{E}\,(\tilde{Y}_{1,t}\,\tilde{W}_{v_i,t})\big)\right)\,, \tag{B.40}$$

$$\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_i,t}^2) = 4\big(\mathrm{E}\,(\tilde{Y}_{1,t}^2\,\tilde{W}_{v_i,t}^2) + \sigma_{v_i}^2\mu_Y^2 - 2\mu_Y\mathrm{E}\,(\tilde{Y}_{1,t}\,\tilde{W}_{v_i,t}^2)\big)\,, \tag{B.41}$$

$$\mathrm{E}\,(\tilde{X}_t^2\,\tilde{W}_{v_1,t}\,\tilde{W}_{v_2,t}) = 4\big(\mathrm{E}\,(\tilde{Y}_{1,t}^2\,\tilde{W}_{v_1,t}\,\tilde{W}_{v_2,t}) + \mu_Y^2\mathrm{E}\,(\tilde{W}_{v_1,t}\,\tilde{W}_{v_2,t}) - 2\mu_Y\mathrm{E}\,(\tilde{Y}_{1,t}\,\tilde{W}_{v_1,t}\,\tilde{W}_{v_2,t})\big)\,. \tag{B.42}$$

For the calculations of $\mathrm{E}\,(\tilde{Y}_{1,t}^p\,\tilde{W}_{v_1,t}^{p_1}\,\tilde{W}_{v_2,t}^{p_2})$, we consider the following cases:

Conditional binomial. This is the case when $\tilde{W}_{v_1,t} = I_{\{P_t>\alpha\}} - (1-\alpha)$.

For the degenerate case when $\beta_1 = \beta_2 = \beta$ for some $\beta \leqslant \alpha$, we obtain $X_t = I_{\{\tilde{P}_t>\beta\}}$, and we can calculate directly

$$\mathrm{E}\,(\tilde{X}_t^p\,\tilde{W}_{v_1,t}^{p_1}) = \beta\,(\beta-1)^p\,(\alpha-1)^{p_1} + (\beta-\alpha)\,\beta^p\,(\alpha-1)^{p_1} + (1-\alpha)\,\beta^p\,\alpha^{p_1}\,. \tag{B.43}$$

Otherwise, assuming that $\gamma_1 < \gamma_2$, we can calculate

$$\begin{aligned}\mathrm{E}\,(\tilde{Y}_{1,t}^p\,\tilde{W}_{v_1,t}^{p_1}) = \gamma_1(-\mu_Y)^p\,(\alpha-1)^{p_1} &+ \int_{\gamma_1}^{\alpha}(u-\gamma_1-\mu_Y)^p(\alpha-1)^{p_1}\,\mathrm{d}u \\ &+ \int_{\alpha}^{\gamma_2}(u-\gamma_1-\mu_Y)^p\,\alpha^{p_1}\,\mathrm{d}u + (1-\gamma_2)(\gamma_2-\gamma_1-\mu_Y)^p\,\alpha^{p_1}\,.\end{aligned} \tag{B.44}$$

Conditional spectral. Assuming that $\gamma_1 < \gamma_2$ and $\alpha_1 < \alpha_2$, and $\mathrm{d}v_1(u) = g_1(u)\,\mathrm{d}u$ with g_1 satisfying Assumption 3.1, we can calculate

$$\begin{aligned}\mathrm{E}\,(\tilde{Y}_{1,t}^p\,\tilde{W}_{v_1,t}^{p_1}) = \gamma_1(-\mu_Y)^p\,(-\mu_{v_1})^{p_1} &+ \int_{\gamma_1}^{\alpha_1}(u-\gamma_1-\mu_Y)^p(-\mu_{v_1})^{p_1}\,\mathrm{d}u \\ &+ \int_{\alpha_1}^{\alpha_2}(u-\gamma_1-\mu_Y)^p(G_1(u)-G_1(\alpha_1)-\mu_{v_1})^{p_1}\,\mathrm{d}u \\ &+ \int_{\alpha_2}^{\gamma_2}(u-\gamma_1-\mu_Y)^p(G_1(\alpha_2)-G_1(\alpha_1)-\mu_{v_1})^{p_1}\,\mathrm{d}u \\ &+ (1-\gamma_2)(\gamma_2-\gamma_1-\mu_Y)^p\,(G_1(\alpha_2)-G_1(\alpha_1)-\mu_{v_1})^{p_1}\,.\end{aligned} \tag{B.45}$$

Conditional bispectral. Assuming that $\gamma_1 < \gamma_2$ and $\alpha_1 < \alpha_2$, and $\mathrm{d}v_i(u) = g_i(u)\,\mathrm{d}u$ with g_i satisfying Assumption 3.1 for $i = 1, 2$, we can calculate

$$\begin{aligned}
\mathrm{E}\,(\tilde{Y}_{1,t}^{p}\,\tilde{W}_{v_1,t}^{p_1}\,\tilde{W}_{v_2,t}^{p_2}) = \gamma_1(-\mu_Y)^p \prod_{i=1}^{2}(-\mu_{v_i})^{p_i} + \int_{\gamma_1}^{\alpha_1}(u-\gamma_1-\mu_Y)^p \prod_{i=1}^{2}(-\mu_{v_i})^{p_i}\,\mathrm{d}u \\
+ \int_{\alpha_1}^{\alpha_2}(u-\gamma_1-\mu_Y)^p \prod_{i=1}^{2}(G_i(u)-G_i(\alpha_1)-\mu_{v_i})^{p_i}\,\mathrm{d}u \\
+ \int_{\alpha_2}^{\gamma_2}(u-\gamma_1-\mu_Y)^p \prod_{i=1}^{2}(G_i(\alpha_2)-G_i(\alpha_1)-\mu_{v_i})^{p_i}\,\mathrm{d}u \\
+ (1-\gamma_2)(\gamma_2-\gamma_1-\mu_Y)^p \prod_{i=1}^{2}(G_i(\alpha_2)-G_i(\alpha_1)-\mu_{v_i})^{p_i}\,.
\end{aligned} \tag{B.46}$$

Conditional PNS. We denote by $\boldsymbol{S}_t(\boldsymbol{\theta}_0)_u = (S_{1,t}(\boldsymbol{\theta}_0)_u, S_{2,t}(\boldsymbol{\theta}_0)_u)^T$ the score vector in (3.78) evaluated at $P_t = u$. Assuming that $\gamma_1 < \gamma_2$ and $\alpha_1 < \alpha_2$, and v_1 and v_2 are the PNS weight measures in (3.80) and (3.81), we can calculate

$$\begin{aligned}
\mathrm{E}\,(\tilde{Y}_{1,t}^{p}\,\tilde{W}_{v_1,t}^{p_1}\,\tilde{W}_{v_2,t}^{p_2}) = \gamma_1(-\mu_Y)^p \prod_{i=1}^{2}(-\mu_{v_i})^{p_i} + \int_{\gamma_1}^{\alpha_1}(u-\gamma_1-\mu_Y)^p \prod_{i=1}^{2}(-\mu_{v_i})^{p_i}\,\mathrm{d}u \\
+ \int_{\alpha_1}^{\alpha_2}(u-\gamma_1-\mu_Y)^p \prod_{i=1}^{2}(S_{i,t}(\boldsymbol{\theta}_0)_u)^{p_i}\,\mathrm{d}u \\
+ \int_{\alpha_2}^{\gamma_2}(u-\gamma_1-\mu_Y)^p \prod_{i=1}^{2}(S_{i,t}(\boldsymbol{\theta}_0)_{\alpha_2})^{p_i}\,\mathrm{d}u \\
+ (1-\gamma_2)(\gamma_2-\gamma_1-\mu_Y)^p \prod_{i=1}^{2}(S_{i,t}(\boldsymbol{\theta}_0)_{\alpha_2})^{p_i}\,.
\end{aligned} \tag{B.47}$$

B.3 Conditional spectral and bispectral test as proposed in Section 5.2.5

For notation simplicity, we define $X_t = f^*(\tilde{P}_t)$ and $\tilde{X}_t = X_t - \hat{\mu}_X$ as in (5.29), where $\hat{\mu}_X$ is the sample mean of X_t (note the notation difference compared to Section 5.2.4, where we have defined $\tilde{X}_t = X_t - \mu_X$). The reason that we use the sample mean $\hat{\mu}_X$ rather than μ_X is given in Section 5.2.5. The required equations are similar to those in Appendix B.1, except that we replace $\tilde{L}_{k,t}$ with

$$\tilde{L}_{k,t} = \tilde{X}_t \, \tilde{X}_{t-k} \,, \tag{B.48}$$

and change the following equations accordingly:

$$\begin{aligned}
(n-h)^4 \, \mathrm{cov}(\overline{W}^2_{v_1}, \overline{L}^2_k) &= \big(2(n-h-k)+k\big)\big(\mathrm{E}\,(\tilde{X}_t^2 \, \tilde{W}^2_{v_1,t})\hat{\sigma}_X^2 - \sigma^2_{v_1}\hat{\sigma}_X^4\big) \\
&\quad + 2(n-h-k)\mathrm{E}\,(\tilde{X}_t^2 \, \tilde{W}_{v_1,t})^2 \\
&\quad + 4 \, \max(n-h-2k,0)\,\hat{\sigma}_X^2 \mathrm{E}\,(\tilde{X}_t \, \tilde{W}_{v_1,t})^2 \,, \qquad \text{(B.49)} \\
(n-h)^4 \, \mathrm{cov}(\overline{W}^2_{v_2}, \overline{L}^2_k) &= \big(2(n-h-k)+k\big)\big(\mathrm{E}\,(\tilde{X}_t^2 \, \tilde{W}^2_{v_2,t})\hat{\sigma}_X^2 - \sigma^2_{v_2}\hat{\sigma}_X^4\big) \\
&\quad + 2(n-h-k)\mathrm{E}\,(\tilde{X}_t^2 \, \tilde{W}_{v_2,t})^2 \\
&\quad + 4 \, \max(n-h-2k,0)\,\hat{\sigma}_X^2 \mathrm{E}\,(\tilde{X}_t \, \tilde{W}_{v_2,t})^2 \,, \qquad \text{(B.50)} \\
(n-h)^4 \, \mathrm{cov}(\overline{W}_{v_1} \, \overline{W}_{v_2}, \overline{L}^2_k) &= \big(2(n-h-k)+k\big)\big(\mathrm{E}\,(\tilde{X}_t^2 \, \tilde{W}_{v_1,t} \, \tilde{W}_{v_2,t})\hat{\sigma}_X^2 - \sigma_{v_1,v_2}\hat{\sigma}_X^4\big) \\
&\quad + 2(n-h-k)\mathrm{E}\,(\tilde{X}_t^2 \, \tilde{W}_{v_1,t})\mathrm{E}\,(\tilde{X}_t^2 \, \tilde{W}_{v_2,t}) \\
&\quad + 4 \, \max(n-h-2k,0)\,\hat{\sigma}_X^2 \mathrm{E}\,(\tilde{X}_t \, \tilde{W}_{v_1,t})\mathrm{E}\,(\tilde{X}_t \, \tilde{W}_{v_2,t}) \,, \qquad \text{(B.51)} \\
(n-h)^4 \, \mathrm{var}(\overline{L}^2_k) &= (n-h)(\mathrm{E}\,(\tilde{X}_t^4)^2 - \hat{\sigma}_X^8) \\
&\quad + 4\,(n-h-k)\,\mathrm{E}\,(\tilde{X}_t^4)\hat{\sigma}_X^4 \\
&\quad + (2\,(n-h)(n-h-1) - 4\,(n-h-k))\hat{\sigma}_X^8 \\
&\quad + 2\,(n-h-k)(\mathrm{E}\,(\tilde{X}_t^4)\hat{\sigma}_X^4 - \hat{\sigma}_X^8) \,, \qquad \text{(B.52)} \\
(n-h)^4 \, \mathrm{cov}(\overline{L}^2_i, \overline{L}^2_j \mid j > i) &= (6i + 4((n-h)-(i+j)) + 2(j-i))(\mathrm{E}\,(\tilde{X}_t^4)\hat{\sigma}_X^4 - \hat{\sigma}_X^8) \\
&\quad + I_{\{j=2i\}}(n-h-i)\mathrm{E}\,(\tilde{X}_t^3)^2\hat{\sigma}_X^2 \\
&\quad + (n-h-j)\hat{\sigma}_X^8 \,, \qquad \text{(B.53)}
\end{aligned}$$

where $\hat{\sigma}_X^2 = \frac{1}{n}\sum_{t=1}^n \tilde{X}_t^2$ is the sample variance of X_t. To be consistent with our treatment of centering X_t using the sample mean $\hat{\mu}_X$, we estimate

$$\mathrm{E}\,(\tilde{X}_t^p \, \tilde{W}^{p_1}_{v_1,t} \, \tilde{W}^{p_2}_{v_2,t}) = \frac{1}{n}\sum_{t=1}^n \tilde{X}_t^p \, \tilde{W}^{p_1}_{v_1,t} \, \tilde{W}^{p_2}_{v_2,t} \,. \tag{B.54}$$

Using (B.54) will result in tests that are slightly undersized compared to if we were to evaluated $\mathrm{E}\left(\tilde{X}_t^p \tilde{W}_{v_1,t}^{p_1} \tilde{W}_{v_2,t}^{p_2}\right)$ under the assumption that (P_t) are iid uniform, especially when the range size $\beta_2 - \beta_1$ or sample size n is small.